

METHOD AND DEVICE FOR VOICE RECOGNITION, AND COMPUTER- READABLE STORAGE MEDIUM

Publication number: JP2001312293

Publication date: 2001-11-09

Inventor: YAMADA MAKI; HOSHIMI MASAKATSU

Applicant: MATSUSHITA ELECTRIC IND CO LTD

Classification:

- International: G10L15/12; G10L15/08; G10L15/28; G10L15/00;
(IPC1-7): G10L15/12; G10L15/08; G10L15/28

- European:

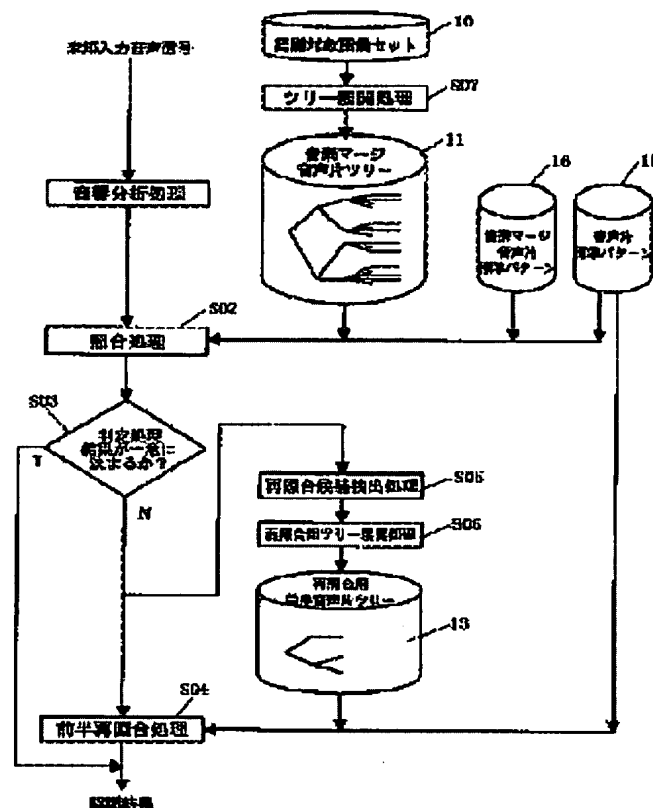
Application number: JP20000129132 20000428

Priority number(s): JP20000129132 20000428

Report a data error here

Abstract of JP2001312293

PROBLEM TO BE SOLVED: To recognize a voice with a small computation quantity without lowering the recognition performance as to a voice recognition technology. **SOLUTION:** This technology has a step wherein a phonetic notation series generated by merging phonemes having similar features of phonetic representation of an object vocabulary set to be recognized is converted into a voice segment series as a minimum unit of recognition and it is expanded into a phoneme merging voice segment tree, a collating step wherein previously found standard patterns representing features of a voice are connected according to the phoneme merging voice segment tree and collated with the feature vector time series of an unknown input voice signal by DP matching using a beam search over time matching, and a step wherein the standard patterns are connected according to a voice segment tree for re-collating if a result is not uniquely determined and then collated with the unknown input voice and the recognition result is outputted. Consequently, a voice can be recognized with a small computation quantity without lowering the recognition performance.



Data supplied from the esp@cenet database - Worldwide

(19)日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11)特許出願公開番号

特開2001-312293

(P2001-312293A)

(43)公開日 平成13年11月9日(2001.11.9)

(51)Int.Cl.

識別記号

F I

テーマコード(参考)

G 1 0 L 15/12

G 1 0 L 3/00

5 3 3 Z 5 D 0 1 5

15/08

5 6 1 J

15/28

5/06

D

審査請求 未請求 請求項の数19 O L (全 23 頁)

(21)出願番号 特願2000-129132(P2000-129132)

(71)出願人 000005821

(22)出願日 平成12年4月28日(2000.4.28)

松下電器産業株式会社

大阪府門真市大字門真1006番地

(72)発明者 山田 麻紀

神奈川県川崎市多摩区東三田3丁目10番1

号 松下技研株式会社内

(72)発明者 星見 昌克

神奈川県川崎市多摩区東三田3丁目10番1

号 松下技研株式会社内

(74)代理人 100097445

弁理士 岩橋 文雄 (外2名)

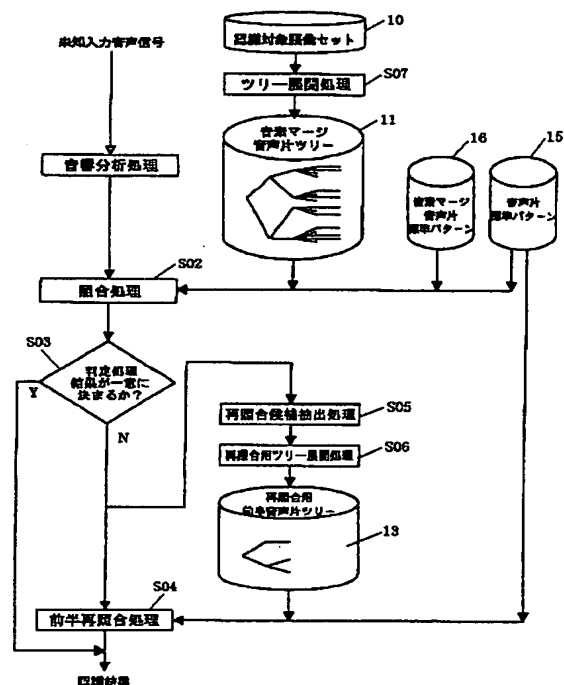
Fターム(参考) 5D015 HH04 HH07 HH11 LL02

(54)【発明の名称】 音声認識方法およびその装置、並びにコンピュータ読み取り可能な記憶媒体

(57)【要約】

【課題】 本発明は音声認識技術に関するものであり、認識性能を落とすことなく少ない計算量で音声の認識を行うことを目的とする。

【解決手段】 認識対象語彙セットの音素表記の特徴の似ている音素をマージした音素表記列を認識の最小単位である音声片列に変換しこれを音素マージ音声片ツリーに展開するステップと、前記音素マージ音声片ツリーに従って、あらかじめ求めておいた音声の特徴を表す標準パターンを接続し、これと未知入力音声信号の特徴ベクトル時系列との照合を、ビームサーチを用いたDPマッチングにより時間整合を取りながら行う照合ステップと、結果が一意に決まらなかった場合に、再照合用音声片ツリーに従って標準パターンを接続し、これと未知入力音声との照合を行い認識結果を出力するステップを有するもので、認識性能を落とすことなく少ない計算量で音声の認識を行うことができる。



【特許請求の範囲】

【請求項1】 未知入力音声信号を音響分析し特徴ベクトル時系列を求めるステップと、認識対象語彙セットの音素表記の特徴の似ている音素をマージした音素表記列を認識の最小単位である音声片列に変換し、これを音素マージ音声片ツリーに展開するステップと、前記音素マージ音声片ツリーに従って、あらかじめ求めておいた音声の特徴を表す音素マージ音声片標準パターンを接続し、これと未知入力音声信号の特徴ベクトル時系列との照合を、ビームサーチを用いたDPマッチングにより時間整合を取りながら行うステップと、照合結果が一意に決まる場合に認識結果を出力するステップと、照合結果が一意に決まらなかった場合に、再照合を行う候補となる認識対象語彙を抽出するステップと、再照合用の音声片ツリーを展開するステップと、再照合用音声片ツリーに従って音声片標準パターンを接続し、これと未知入力音声との照合を、ビームサーチを用いたDPマッチングにより時間整合を取りながら行い認識結果を出力するステップとを有することを特徴とする音声認識方法。

【請求項2】 音素マージ音声片ツリーに展開するステップは、認識対象語彙セットの音素表記の語頭から第N番目の音素までのうち特徴の似ている音素をマージした音素表記列を認識の最小単位である音声片列に変換し、これを音素マージ音声片ツリーに展開することを特徴とする請求項1記載の音声認識方法。

【請求項3】 音素のマージは、子音を音素群毎にまとめてマージすることを特徴とする請求項1または2記載の音声認識方法。

【請求項4】 未知入力音声信号を音響分析し特徴ベクトル時系列を求めるステップと、認識対象語彙セットの音素表記列の語頭から第N番目の音素までを、精度の粗い音声片標準パターンを持つラフ音声片の系列に変換し、第N番目以降の音素を精密な音声片標準パターンを持つ精密音声片の系列に変換し、これをラフ音声片ツリーに展開するステップと、前記ラフ音声片ツリーに従って、あらかじめ求めておいた音声の特徴を表す音声片標準パターンおよび音素マージ音声片標準パターンを接続し、これと未知入力音声信号の特徴ベクトル時系列との照合を、ビームサーチを用いたDPマッチングにより時間整合を取りながら行うステップと、照合の結果から再照合を行う候補となる認識対象語彙を抽出するステップと、再照合用の音声片ツリーを展開するステップと、再照合用音声片ツリーに従って精密な音声片標準パターンを接続し、これと未知入力音声との照合を、ビームサーチを用いたDPマッチングにより時間整合を取りながら行い認識結果を出力するステップとを有することを特徴とする音声認識方法。

【請求項5】 未知入力音声信号を音響分析し特徴ベクトル時系列を求めるステップと、認識対象語彙セットの音素表記列の語頭から第N番目の音素までを、精度の粗

い音声片標準パターンを持つラフ音声片の系列に変換し、第N番目以降の音素を精密な音声片標準パターンを持つ精密音声片の系列に変換し、これをラフ音声片ツリーに展開するステップと、前記ラフ音声片ツリーに従って、あらかじめ求めておいた音声の特徴を表す音声片標準パターンおよび音素マージ音声片標準パターンを接続し、これと未知入力音声信号の特徴ベクトル時系列との照合を、ビームサーチを用いたDPマッチングにより時間整合を取りながら行い認識結果を出力するステップとを有することを特徴とする音声認識方法。

【請求項6】 精度を粗い音声片の標準パターンは、ラフ音声片1つの音声片あたりにかかる距離計算量を精密音声片1つあたりにかかる計算量に比べ削減することを特徴とする請求項4または5記載の音声認識方法。

【請求項7】 精度を粗い音声片の標準パターンは、認識結果が一意に決まる範囲内で、異なる音韻環境の音声片をマージすることを特徴とする請求項4または5記載の音声認識方法。

【請求項8】 音声片の距離計算量にかかるコストの削減は、ラフ音声片標準パターンのフレーム数を少なくすることを特徴とする請求項6記載の音声認識方法。

【請求項9】 音声片の距離計算にかかるコストの削減は、特徴パラメータベクトルの出現確率が複数のガウス分布の和、すなわちガウス分布の混合分布で近似できると仮定したとき、ラフ音声片標準パターンのガウス分布の混合数を少なくすることを特徴とする請求項6記載の音声認識方法。

【請求項10】 音声片の距離計算にかかるコストの削減は、特徴パラメータベクトルの出現確率が複数のガウス分布の和、すなわちガウス分布の混合分布で近似できると仮定したとき、ラフ音声片標準パターンのガウス分布の共分散行列の種類数を少なくすることを特徴とする請求項9記載の音声認識方法。

【請求項11】 再照合の際、未知入力音声の前半部分とのみ照合を行い認識結果を出力することを特徴とする請求項1、4、5のいずれかに記載の音声認識方法。

【請求項12】 再照合の際、未知入力音声の発声区間すべてと照合を行い認識結果を出力することを特徴とする請求項1、4、5のいずれかに記載の音声認識方法。

【請求項13】 未知入力音声の発声区間を特定せず、異なる始端を認める連続DPマッチングを用いたことを特徴とする請求項1、4、5のいずれかに記載の音声認識方法。

【請求項14】 未知入力音声信号を音響分析し特徴ベクトル時系列を求める音響分析手段と、認識対象語彙セットの音素表記の特徴の似ている音素をマージした音素表記列を認識の最小単位である音声片列に変換し、これを音素マージ音声片ツリーに展開するツリー展開手段と、前記音素マージ音声片ツリーに従って、あらかじめ求めておいた音声の特徴を表す音素マージ音声片標準パ

ターンを接続し、これと未知入力音声信号の特徴ベクトル時系列との照合を、ビームサーチを用いたDPマッチングにより時間整合を取りながら行う照合手段と、照合結果が一意に決まるか否かを判定し、一意に決まる場合に認識結果を出力する判定手段と、照合結果が一意に決まらなかった場合に、再照合を行う候補となる認識対象語彙を抽出する再照合候補抽出手段と、再照合用の音声片ツリーを展開する再照合用ツリー展開手段と、再照合用音声片ツリーに従って音声片標準パターンを接続し、これと未知入力音声との照合を、ビームサーチを用いたDPマッチングにより時間整合を取りながら行い認識結果を出力する再照合手段とを有することを特徴とする音声認識装置。

【請求項15】 未知入力音声信号を音響分析し特徴ベクトル時系列を求める音響分析手段と、認識対象語彙セットの音素表記列の語頭から第N番目の音素までを、精度の粗い音声片標準パターンを持つラフ音声片の系列に変換し、第N番目以降の音素を精密な音声片標準パターンを持つ精密音声片の系列に変換し、これをラフ音声片ツリーに展開するツリー展開手段と、前記ラフ音声片ツリーに従って、あらかじめ求めておいた音声の特徴を表す音声片標準パターンおよび音素マージ音声片標準パターンを接続し、これと未知入力音声信号の特徴ベクトル時系列との照合を、ビームサーチを用いたDPマッチングにより時間整合を取りながら行う照合手段と、照合の結果から再照合を行う候補となる認識対象語彙を抽出する再照合候補抽出手段と、再照合用の音声片ツリーを展開する再照合ツリー展開手段と、再照合用音声片ツリーに従って精密な音声片標準パターンを接続し、これと未知入力音声との照合を、ビームサーチを用いたDPマッ

チングにより時間整合を取りながら行い認識結果を出力する再照合手段とを有することを特徴とする音声認識装置。

【請求項16】 未知入力音声信号を音響分析し特徴ベクトル時系列を求める音響分析手段と、認識対象語彙セットの音素表記列の語頭から第N番目の音素までを、精度の粗い音声片標準パターンを持つラフ音声片の系列に変換し、第N番目以降の音素を精密な音声片標準パターンを持つ精密音声片の系列に変換し、これをラフ音声片ツリーに展開するツリー展開手段と、前記ラフ音声片ツリーに従って、あらかじめ求めておいた音声の特徴を表す音声片標準パターンおよび音素マージ音声片標準パターンを接続し、これと未知入力音声信号の特徴ベクトル時系列との照合を、ビームサーチを用いたDPマッチングにより時間整合を取りながら行い認識結果を出力する照合手段とを有することを特徴とする音声認識装置。

【請求項17】 プログラムされたコンピュータによって音声を認識するプログラムを記録した記録媒体であって、未知入力音声信号を音響分析し特徴ベクトル時系列を求めるステップと、認識対象語彙セットの音素表記の

特徴の似ている音素をマージした音素表記列を認識の最小単位である音声片列に変換し、これを音素マージ音声片ツリーに展開するステップと、前記音素マージ音声片ツリーに従って、あらかじめ求めておいた音声の特徴を表す音素マージ音声片標準パターンを接続し、これと未知入力音声信号の特徴ベクトル時系列との照合を、ビームサーチを用いたDPマッチングにより時間整合を取りながら行うステップと、照合結果が一意に決まる場合に認識結果を出力するステップと、照合結果が一意に決まらなかった場合に、再照合を行う候補となる認識対象語彙を抽出するステップと、再照合用の音声片ツリーを展開するステップと、再照合用音声片ツリーに従って音声片標準パターンを接続し、これと未知入力音声との照合を、ビームサーチを用いたDPマッチングにより時間整合を取りながら行い認識結果を出力するステップとを有することを特徴とするコンピュータ読み取り可能な記憶媒体。

【請求項18】 プログラムされたコンピュータによって音声を認識するプログラムを記録した記録媒体であって、未知入力音声信号を音響分析し特徴ベクトル時系列を求めるステップと、認識対象語彙セットの音素表記列の語頭から第N番目の音素までを、精度の粗い音声片標準パターンを持つラフ音声片の系列に変換し、第N番目以降の音素を精密な音声片標準パターンを持つ精密音声片の系列に変換し、これをラフ音声片ツリーに展開するステップと、前記ラフ音声片ツリーに従って、あらかじめ求めておいた音声の特徴を表す音声片標準パターンおよび音素マージ音声片標準パターンを接続し、これと未知入力音声信号の特徴ベクトル時系列との照合を、ビームサーチを用いたDPマッチングにより時間整合を取りながら行うステップと、照合の結果から再照合を行う候補となる認識対象語彙を抽出するステップと、再照合用の音声片ツリーを展開するステップと、再照合用音声片ツリーに従って精密な音声片標準パターンを接続し、これと未知入力音声との照合を、ビームサーチを用いたDPマッチングにより時間整合を取りながら行い認識結果を出力するステップとを有することを特徴とするコンピュータ読み取り可能な記憶媒体。

【請求項19】 プログラムされたコンピュータによって音声を認識するプログラムを記録した記録媒体であって、未知入力音声信号を音響分析し特徴ベクトル時系列を求めるステップと、認識対象語彙セットの音素表記列の語頭から第N番目の音素までを、精度の粗い音声片標準パターンを持つラフ音声片の系列に変換し、第N番目以降の音素を精密な音声片標準パターンを持つ精密音声片の系列に変換し、これをラフ音声片ツリーに展開するステップと、前記ラフ音声片ツリーに従って、あらかじめ求めておいた音声の特徴を表す音声片標準パターンおよび音素マージ音声片標準パターンを接続し、これと未知入力音声信号の特徴ベクトル時系列との照合を、ビー

ムサーチを用いたDPマッチングにより時間整合を取りながら行い認識結果を出力するステップとを有することを特徴とするコンピュータ読み取り可能な記憶媒体。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、ビームサーチを用いたDPマッチングを用いた音声認識方法およびその装置、並びにコンピュータ読み取り可能な記憶媒体に関するものである。

【0002】

【従来の技術】認識対象となる音声の特徴を表現した標準パターンと、未知入力音声信号の特徴ベクトル時系列との照合を、ビームサーチを用いたDPマッチングにより時間整合を取りながら行い認識する音声認識方法として、日本音響学会講演論文集、平成9年9月、3-1-4「音素の特徴点に着目した大語彙不特定話者音声認識法」が知られている。

【0003】図16は、従来の音声認識装置のコンピュータを用いた構成図を示すものである。図16において、1は音声を取り込むマイク、2はA/D、3はインタフェース(I/F)、4はメモリ、5はCPU、6はキーボード/ディスプレイ、7はCPUバス、8はI/F、9は出力、10は認識対象語彙セット、15は音声片標準パターン、19は音声片ツリーである。

【0004】上記のように構成された従来の音声認識装置の動作を図17のフローチャートを用いて説明する。まず最初に、従来例における認識辞書にあたる音声片ツリー19について、図3、図4を参照しながら説明をする。

【0005】標準パターンの単位として、音素片、音素、音節、CV/VC(子音+母音/母音+子音)、V CV、CVCなどが考えられる。これら認識の最小単位を音声片と呼ぶ。従来例では、子音の始端から母音中心までを表すCVと、母音中心から母音終端までを表すVC、母音中心から母音中心までを表すVVを基本の単位とする。VCは母音区間しか含まないが、後続子音により異なるVCと定義する。

【0006】たとえば、認識対象語彙を「きりはら」「きりゅう」「ちり」「ちりゅう」「めぐろ」「めむろ」「ねむろ」「ふちゅう」の8単語としたとき、これらを音声片列で表すと、図4のようになる。

【0007】これを単純なツリー構造で表したものが図3である。これを音声片ツリーと定義する。ここでは、アークに音声片を割り当てたが、ノードに割り当てるともできる。語彙の終端にあたるノードには、その語彙の終端であることがわかるようにしておく。このような

ノードをリーフノードと定義する。図3ではリーフノードを黒丸で表している。また、ツリーの深さを、根から数えて第1段、第2段、…と数えるとする。

【0008】以下、従来例について、図17のフローチャートを参照しながらその動作を説明する。

【0009】音声片標準パターン15は、あらかじめ多数話者が発声した学習データから学習し、音声片毎に求めておく。本従来例では、特徴パラメータベクトルの出現確率が複数のガウス分布の和(これを混合分布と呼ぶ)で近似できると仮定し、学習データから、標準パターンのフレームごとにガウス分布の平均値ベクトルおよび共分散行列を求め、これを音声片標準パターン15とする。

【0010】音声片ツリー19は、あらかじめ認識対象語彙セット10から、ツリー展開処理S06において作成しておく。

【0011】まず、音響分析処理S01は、入力された未知音声信号を分析時間(以下フレームと呼ぶ)毎にN個の特徴パラメータに変換される。特徴パラメータとしては、線形予測分析によるLPCケプストラム係数、LPCメルケプストラム係数、メル線形予測分析によるメルLPCケプストラム係数、メルスケールフィルタバンクによるメル周波数ケプストラム係数(MFCC)など、音声認識に適したものならばどのようなものを用いても良い。

【0012】照合処理S02では、音声片ツリー19にしたがって音声片標準パターン15を接続しながら、上記未知入力音声の特徴パラメータ時系列と標準パターンとの照合を行う。照合は、入力フレーム同期のビームサーチを用いたDPマッチングにより行う。照合の結果最も累積スコアの高かったリーフノードを求め、このリーフノードに対応する語彙を認識結果として出力する。

【0013】以下に、DPマッチングによる照合と、ビームサーチによる枝刈りのアルゴリズムについて説明する。

【0014】DPマッチングは、入力音声と標準パターンの時間整合をとりながら照合する方法である。第j番目の入力フレームと、音声片ツリーの第k番目のアークに対応する音声片mの標準パターンの第i番目のフレームとの累積スコア $L(i, k; j)$ は、次の漸化式で表される。ただし $d(i, m; j)$ は入力の第jフレームと音声片mの標準パターンの第iフレームとの距離である。

【0015】

【数1】

7
(初期化)
 $L(0,0;0) = 0$

(for $j=1, \dots, J$)

(for $k=1, \dots, K$)

$L(0, k; j-1) = L(i_k^k, k; j-1)$
 i_k^k は、アーク k の前に接続するアーク番号、 i_k^k はアーク k の標準パターンの最終フレームを表す。

(for $i=1, \dots, I_k$)

$L(i, k; j) = \max \begin{cases} L(i, k; j-1) + d(i, m; j) \\ L(i-1, k; j-1) + d(i, m; j) \end{cases}$

【0016】発声終了時に、リーフノードの累積スコア（終端アークの終端フレームにおける累積スコア）で最も大きいものが認識結果のスコアとなる。

【0017】ビームサーチは、DPマッチングの際スコアの低い経路は計算せずに、スコアの高い経路だけを伸張させながら計算していく手法である。累積スコアの低い経路であるかどうかは、入力と辞書の格子点における累積スコアの値が、その1フレーム前の時刻の最大累積スコアに比べ一定値（ビーム幅）以上低くなっているかどうかで判定する。累積スコアの低い格子点を枝刈りし、それ以外の格子点を候補として残していく。以下にそのアルゴリズムを示す。

【0018】入力フレーム同期に、以下の式にしたがって、格子点 $(i, k; j)$ を枝刈るか、候補として残すかの次の式によって判定をしながらDPパスを伸ばしていくものである。

【0019】

【数2】

$$\theta = \max_{i,k} \{L(i, k; j-1)\} - \alpha$$

$L(i, k; j) \geq \theta$ のとき、その格子点は残す。

$L(i, k; j) < \theta$ のとき、その格子点は枝刈る。

ただし、 α はビーム幅

【0020】ビームサーチを用いたDPマッチングでは、発声開始付近ではまだどの候補も大きなスコアの差がつかないため、枝刈りはあまり行われない。そして発声後しばらくすると、発声内容とかけ離れた候補が枝刈りされはじめる。

【0021】一方、認識対象語彙数が多い場合、語頭付近の音声片の種類数は非常に多くなる。そのため上記従来の構成では、音声片ツリーは第1段目から大きく広がってしまい、発声開始付近では、照合のために非常に多くの経路について計算しなくてはならなくなってしまう。これはすなわち、発声開始付近では探索空間が広い

と言える。

【0022】したがって、発声開始付近では、探索空間が広い上に枝刈りがあまりなされないために、格子点候補数は爆発的に増えてしまう。発声開始からしばらくすると、探索空間は広くても、枝刈りが多くなされるようになるため、格子点候補数は急激に減少する。

【0023】従来法では、格子点候補数に比例して、認識にかかる計算量も増大する。したがって、従来法では認識にかかる計算量は図18のように時間変化する。図18を見てわかるとおり、発声開始付近での計算量は極端に多くなり、全体の計算量を削減するためには、発声開始付近での計算量を削減することが肝要である。

【0024】単純に発声開始付近でのビーム幅を狭めることによって発声開始付近での計算量を削減することはできるが、その場合正解候補が枝刈られやすくなる。発声開始付近では発声の言いよどみなどが起こりやすく、語頭のスコアが悪いというだけで枝刈りをしてしまうのは問題である。

【0025】

【発明が解決しようとする課題】しかしながら上記の従来の構成では、発声開始付近で計算量が極端に多くなるという課題を有していた。

【0026】本発明は、上記従来の課題を解決するもので、正解パスが枝刈られないようにしつつ語頭付近の探索空間を小さくするまたは語頭付近での照合にかかる計算量を削減する、すなわち認識性能を落とさずに全体の計算量を削減することを目的とする。

【0027】

【課題を解決するための手段】この課題を解決するために、本発明は、音声片ツリーの語頭付近の広がりを含めた音素マージ音声片ツリーを用いることによって格子点候補数を削減する、または音声片ツリーの語頭付近における音声片標準パターンの精度を粗くしたラフ音声片ツリーを用いることによって照合にかかる計算量を削減す

る。

【0028】これは、ビームサーチで計算量の多い発声の前半部分は粗い照合を、ビームサーチで計算量の少なくなる後半部分は精密な照合をするという考え方に基づくものである。

【0029】これにより、発声開始付近での計算量が削減し、認識性能を落とさずに計算量を削減することができる。

【0030】

【発明の実施の形態】本発明の請求項1に記載の発明は、未知入力音声信号を音響分析し特徴ベクトル時系列を求めるステップと、認識対象語彙セットの音素表記の特徴の似ている音素をマージした音素表記列を認識の最小単位である音声片列に変換し、これを音素マージ音声片ツリーに展開するステップと、前記音素マージ音声片ツリーに従って、あらかじめ求めておいた音声の特徴を表す音素マージ音声片標準パターンを接続し、これと未知入力音声信号の特徴ベクトル時系列との照合を、ビームサーチを用いたDPマッチングにより時間整合を取りながら行うステップと、照合結果が一意に決まる場合に認識結果を出力するステップと、照合結果が一意に決まらなかった場合に、再照合を行う候補となる認識対象語彙を抽出するステップと、再照合用の音声片ツリーを展開するステップと、再照合用音声片ツリーに従って音声片標準パターンを接続し、これと未知入力音声との照合を、ビームサーチを用いたDPマッチングにより時間整合を取りながら行い認識結果を出力するステップとを有するものであり、音素をマージすることにより音声片ツリーの広がり小さくなるため、探索空間が小さくなり1回目の照合にかかる計算量を大幅に削減することができる、再照合を行ったとしても全体の計算量を削減できるという作用を有する。さらに1回目の照合では音素マージを行うことにより特徴の似ている語彙は区別せずに認識を行うため1回目の照合で正解候補が漏れる可能性が低いという利点がある。

【0031】請求項2に記載の発明は、請求項1記載の音声認識方法において、音素マージ音声片ツリーに展開するステップは、認識対象語彙セットの音素表記の語頭から第N番目の音素までのうち特徴の似ている音素をマージした音素表記列を認識の最小単位である音声片列に変換し、これを音素マージ音声片ツリーに展開することを特徴とするものであり、語頭から第N番目の音素までのうち特徴の似ている音素をマージすることにより、特に探索空間の広い語頭付近のツリーの広がりを狭めることができるため、探索空間が小さくなり1回目の照合にかかる計算量を大幅に削減することができ、再照合を行ったとしても全体の計算量を削減できるという作用を有する。さらに1回目の照合では音素マージを行うことにより特徴の似ている語彙は区別せずに認識を行うため1回目の照合で正解候補が漏れる可能性が低いという利点

がある。

【0032】請求項3に記載の発明は、請求項1または2記載の音声認識方法において、音素のマージは、子音を音素群毎にまとめてマージすることを特徴とするものであり、カテゴリ数が多く比較的認識が難しい子音を音響特徴の似通った音素群毎にまとめてマージするため、マージによる誤差が小さく認識性能を落とさずに効率よく計算量削減することができるという作用を有する。

【0033】請求項4に記載の発明は、未知入力音声信号を音響分析し特徴ベクトル時系列を求めるステップと、認識対象語彙セットの音素表記列の語頭から第N番目の音素までを、精度の粗い音声片標準パターンを持つラフ音声片の系列に変換し、第N番目以降の音素を精密な音声片標準パターンを持つ精密音声片の系列に変換し、これをラフ音声片ツリーに展開するステップと、前記ラフ音声片ツリーに従って、あらかじめ求めておいた音声の特徴を表す音声片標準パターンおよび音素マージ音声片標準パターンを接続し、これと未知入力音声信号の特徴ベクトル時系列との照合を、ビームサーチを用いたDPマッチングにより時間整合を取りながら行うステップと、照合の結果から再照合を行う候補となる認識対象語彙を抽出するステップと、再照合用の音声片ツリーを展開するステップと、再照合用音声片ツリーに従って精密な音声片標準パターンを接続し、これと未知入力音声との照合を、ビームサーチを用いたDPマッチングにより時間整合を取りながら行い認識結果を出力するステップとを有するものであり、語頭付近での1回目の照合では音声片ツリーの前半部分は精度の粗いラフ音声片標準パターンを用いるため、1回目の照合にかかる計算量を大幅に削減することができ、再照合を行っても全体の計算量を削減できるという作用を有する。また再照合を行うことにより認識性能を落とさずに認識することができる。

【0034】請求項5に記載の発明は、未知入力音声信号を音響分析し特徴ベクトル時系列を求めるステップと、認識対象語彙セットの音素表記列の語頭から第N番目の音素までを、精度の粗い音声片標準パターンを持つラフ音声片の系列に変換し、第N番目以降の音素を精密な音声片標準パターンを持つ精密音声片の系列に変換し、これをラフ音声片ツリーに展開するステップと、前記ラフ音声片ツリーに従って、あらかじめ求めておいた音声の特徴を表す音声片標準パターンおよび音素マージ音声片標準パターンを接続し、これと未知入力音声信号の特徴ベクトル時系列との照合を、ビームサーチを用いたDPマッチングにより時間整合を取りながら行い認識結果を出力するステップとを有するものであり、音声片ツリーの前半部分のみ精度の粗いラフ音声片標準パターンを用いて照合し、再照合をしないため、計算量は大幅に削減できるという作用を有する。音声片の標準パター

ンの精度を粗くするのは探索空間の広い語頭付近だけであるため、一律に音声片の標準パターンの精度を粗くするよりも効率的に計算量削減することができる。

【0035】請求項6に記載の発明は、請求項4または5記載の音声認識方法において、精度を粗い音声片の標準パターンは、ラフ音声片1つの音声片あたりにかかる距離計算量を精密音声片1つあたりにかかる計算量に比べ削減することを特徴とするものであり、1つの音声片あたりにかかる距離計算量を削減する処理を設けることにより、容易に計算量を削減することができるという作用を有する。

【0036】請求項7に記載の発明は、請求項4または5記載の音声認識方法において、精度を粗い音声片の標準パターンは、認識結果が一意に決まる範囲内で、異なる音韻環境の音声片をマージすることを特徴とするものであり、認識結果が一意に決まる範囲内で、異なる音韻環境の音声片をマージする処理を設けることにより、語頭付近の探索空間が狭まり効率よく計算量を削減することができるという作用を有する。

【0037】請求項8に記載の発明は、請求項6記載の音声認識方法において、音声片の距離計算量にかかるコストの削減は、ラフ音声片標準パターンのフレーム数を少なくすることを特徴とするものであり、音声片標準パターンのフレーム数を削減する処理を設けることにより、容易に計算量を削減することができるという作用を有する。

【0038】請求項9に記載の発明は、請求項6記載の音声認識方法において、音声片の距離計算にかかるコストの削減は、特徴パラメータベクトルの出現確率が複数のガウス分布の和、すなわちガウス分布の混合分布で近似できると仮定したとき、ラフ音声片標準パターンのガウス分布の混合数を少なくすることを特徴とするものであり、音声片標準パターンのガウス分布の混合数を削減する処理を設けることにより、容易に計算量を削減することができるという作用を有する。

【0039】請求項10に記載の発明は、請求項9記載の音声認識方法において、音声片の距離計算にかかるコストの削減は、特徴パラメータベクトルの出現確率が複数のガウス分布の和、すなわちガウス分布の混合分布で近似できると仮定したとき、ラフ音声片標準パターンのガウス分布の共分散行列の種類数を少なくすることを特徴とするものであり、音声片標準パターンのガウス分布の共分散行列を共通化する処理を設けることにより、容易に計算量を削減することができるという作用を有する。

【0040】請求項11に記載の発明は、請求項1、4、5のいずれかに記載の音声認識方法において、再照合の際、未知入力音声の前半部分とのみ照合を行い認識結果を出力することを特徴とするものであり、未知入力音声の前半部分とのみ照合を行い認識結果を出力する処

理を設けることにより、再照合する区間が短くてすむため、再照合にかかる計算量を抑えることができるという作用を有する。

【0041】請求項12に記載の発明は、請求項1、4、5のいずれかに記載の音声認識方法において、再照合の際、未知入力音声の発声区間すべてと照合を行い認識結果を出力することを特徴とするものであり、未知入力音声の発声区間すべてと照合を行い認識結果を出力する処理を設けることにより、より精密な再照合が行えるため認識性能の劣化が少なくすむという作用を有する。

【0042】請求項13に記載の発明は、請求項1、4、5のいずれかに記載の音声認識方法において、未知入力音声の発声区間を特定せず、異なる始端を認める連続DPマッチングを用いたことを特徴とするものであり、発声区間を特定しなくても、認識することができるという作用を有する。

【0043】請求項14に記載の発明は、未知入力音声信号を音響分析し特徴ベクトル時系列を求める音響分析手段と、認識対象語彙セットの音素表記の特徴の似ている音素をマージした音素表記列を認識の最小単位である音声片列に変換し、これを音素マージ音声片ツリーに展開するツリー展開手段と、前記音素マージ音声片ツリーに従って、あらかじめ求めておいた音声の特徴を表す音素マージ音声片標準パターンを接続し、これと未知入力音声信号の特徴ベクトル時系列との照合を、ビームサーチを用いたDPマッチングにより時間整合を取りながら行う照合手段と、照合結果が一意に決まるか否かを判定し、一意に決まる場合に認識結果を出力する判定手段と、照合結果が一意に決まらなかった場合に、再照合を行う候補となる認識対象語彙を抽出する再照合候補抽出手段と、再照合用の音声片ツリーを展開する再照合用ツリー展開手段と、再照合用音声片ツリーに従って音声片標準パターンを接続し、これと未知入力音声との照合を、ビームサーチを用いたDPマッチングにより時間整合を取りながら行い認識結果を出力する再照合手段とを有するもので、音素をマージすることにより音声片ツリーの広がり小さくなるため、探索空間が小さくなり1回目の照合にかかる計算量を大幅に削減することができる、再照合を行ったとしても全体の計算量を削減できるという作用を有する。

【0044】請求項15に記載の発明は、未知入力音声信号を音響分析し特徴ベクトル時系列を求める音響分析手段と、認識対象語彙セットの音素表記列の語頭から第N番目の音素までを、精度の粗い音声片標準パターンを持つラフ音声片の系列に変換し、第N番目以降の音素を精密な音声片標準パターンを持つ精密音声片の系列に変換し、これをラフ音声片ツリーに展開するツリー展開手段と、前記ラフ音声片ツリーに従って、あらかじめ求めておいた音声の特徴を表す音声片標準パターンおよび音

素マージ音声片標準パターンを接続し、これと未知入力音声信号の特徴ベクトル時系列との照合を、ビームサーチを用いたDPマッチングにより時間整合を取りながら行う照合手段と、照合の結果から再照合を行う候補となる認識対象語彙を抽出する再照合候補抽出手段と、再照合用の音声片ツリーを展開する再照合ツリー展開手段と、再照合用音声片ツリーに従って精密な音声片標準パターンを接続し、これと未知入力音声との照合を、ビームサーチを用いたDPマッチングにより時間整合を取りながら行い認識結果を出力する再照合手段とを有するもので、語頭付近での1回目の照合では音声片ツリーの前半部分は精度の粗いラフ音声片標準パターンを用いるため、1回目の照合にかかる計算量を大幅に削減することができ、再照合を行っても全体の計算量を削減できるという作用を有する。

【0045】請求項16に記載の発明は、未知入力音声信号を音響分析し特徴ベクトル時系列を求める音響分析手段と、認識対象語彙セットの音素表記列の語頭から第N番目の音素までを、精度の粗い音声片標準パターンを持つラフ音声片の系列に変換し、第N番目以降の音素を精密な音声片標準パターンを持つ精密音声片の系列に変換し、これをラフ音声片ツリーに展開するツリー展開手段と、前記ラフ音声片ツリーに従って、あらかじめ求めておいた音声の特徴を表す音声片標準パターンおよび音素マージ音声片標準パターンを接続し、これと未知入力音声信号の特徴ベクトル時系列との照合を、ビームサーチを用いたDPマッチングにより時間整合を取りながら行い認識結果を出力する照合手段とを有するもので、音声片ツリーの前半部分のみ精度の粗いラフ音声片標準パターンを用いて照合し、再照合をしないため、計算量は大幅に削減できるという作用を有する。

【0046】請求項17に記載の発明は、プログラムされたコンピュータによって音声認識するプログラムを記録した記録媒体であって、未知入力音声信号を音響分析し特徴ベクトル時系列を求めるステップと、認識対象語彙セットの音素表記の特徴の似ている音素をマージした音素表記列を認識の最小単位である音声片列に変換し、これを音素マージ音声片ツリーに展開するステップと、前記音素マージ音声片ツリーに従って、あらかじめ求めておいた音声の特徴を表す音素マージ音声片標準パターンを接続し、これと未知入力音声信号の特徴ベクトル時系列との照合を、ビームサーチを用いたDPマッチングにより時間整合を取りながら行うステップと、照合結果が一意に決まる場合に認識結果を出力するステップと、照合結果が一意に決まらなかった場合に、再照合を行う候補となる認識対象語彙を抽出するステップと、再照合用の音声片ツリーを展開するステップと、再照合用音声片ツリーに従って音声片標準パターンを接続し、これと未知入力音声との照合を、ビームサーチを用いたDPマッチングにより時間整合を取りながら行い認識結果

を出力するステップとを有することを特徴とするコンピュータ読み取り可能な記憶媒体であり、コンピュータに読み込み実行するものであり、音素をマージすることにより音声片ツリーの広がり小さくなるため、探索空間が小さくなり1回目の照合にかかる計算量を大幅に削減することができ、再照合を行っても全体の計算量を削減できるという作用を有する。

【0047】請求項18に記載の発明は、プログラムされたコンピュータによって音声認識するプログラムを記録した記録媒体であって、未知入力音声信号を音響分析し特徴ベクトル時系列を求めるステップと、認識対象語彙セットの音素表記列の語頭から第N番目の音素までを、精度の粗い音声片標準パターンを持つラフ音声片の系列に変換し、第N番目以降の音素を精密な音声片標準パターンを持つ精密音声片の系列に変換し、これをラフ音声片ツリーに展開するステップと、前記ラフ音声片ツリーに従って、あらかじめ求めておいた音声の特徴を表す音声片標準パターンおよび音素マージ音声片標準パターンを接続し、これと未知入力音声信号の特徴ベクトル時系列との照合を、ビームサーチを用いたDPマッチングにより時間整合を取りながら行うステップと、照合の結果から再照合を行う候補となる認識対象語彙を抽出するステップと、再照合用の音声片ツリーを展開するステップと、再照合用音声片ツリーに従って精密な音声片標準パターンを接続し、これと未知入力音声との照合を、ビームサーチを用いたDPマッチングにより時間整合を取りながら行い認識結果を出力するステップとを有することを特徴とするコンピュータ読み取り可能な記憶媒体であり、コンピュータに読み込み実行するもので、語頭付近での1回目の照合では音声片ツリーの前半部分は精度の粗いラフ音声片標準パターンを用いるため、1回目の照合にかかる計算量を大幅に削減することができ、再照合を行っても全体の計算量を削減できるという作用を有する。

【0048】請求項19に記載の発明は、プログラムされたコンピュータによって音声認識するプログラムを記録した記録媒体であって、未知入力音声信号を音響分析し特徴ベクトル時系列を求めるステップと、認識対象語彙セットの音素表記列の語頭から第N番目の音素までを、精度の粗い音声片標準パターンを持つラフ音声片の系列に変換し、第N番目以降の音素を精密な音声片標準パターンを持つ精密音声片の系列に変換し、これをラフ音声片ツリーに展開するステップと、前記ラフ音声片ツリーに従って、あらかじめ求めておいた音声の特徴を表す音声片標準パターンおよび音素マージ音声片標準パターンを接続し、これと未知入力音声信号の特徴ベクトル時系列との照合を、ビームサーチを用いたDPマッチングにより時間整合を取りながら行い認識結果を出力するステップとを有することを特徴とするコンピュータ読み取り可能な記憶媒体であり、コンピュータに読み込み実

行するもので、音声片ツリーの前半部分のみ精度の粗いラフ音声片標準パターンを用いて照合し、再照合をしないため、計算量は大幅に削減できるという作用を有する。

【0049】以下、本発明の実施の形態について図を用いて説明する。

【0050】（実施の形態1）図1は、本発明の実施の形態1における音声認識装置のブロック構成図であり、以下に説明する。

【0051】図1において、1は音声を取り込むマイク、2はA/D、3はインタフェース（I/F）、4はメモリ、5はCPU、6はキーボード/ディスプレイ、7はCPUバス、8はI/F、9は出力、10は認識対象語彙セット、11は音素マージ音声片ツリー、12はラフ音声片ツリー、13は再照合用前半音声片ツリー、14は再照合用音声片ツリー、15は音声片標準パターン、16は音素マージ音声片標準パターン、17はラフ音声片標準パターン、18は精密音声片標準パターンである。

【0052】まず最初に、実施の形態1における認識辞書に当たる音素マージ音声片ツリー11について、図3、図4、図5を参照しながら説明をする。

【0053】標準パターンの単位として、音素片、音素、音節、CV/VC（子音+母音/母音+子音）、V CV、CVCなどが考えられる。これら認識の最小単位を音声片と呼ぶ。本実施の形態では、子音の始端から母音中心までを表すCVと、母音中心から母音終端までを表すVC、母音中心から母音中心までを表すVVを基本の単位とする。VCは母音区間しか含まないが、後続子音により異なるVCと定義する。

【0054】たとえば、認識対象語彙を「きりはら」「きりゅう」「ちり」「ちりゅう」「めぐろ」「めむろ」「ねむろ」「ふちゅう」の8単語としたとき、これらを音声片列で表すと、図4のようになる。

【0055】これを単純なツリー構造で表したものが図3である。本実施の形態では、これを基本音声片ツリーと定義する。これは従来例で用いている音声片ツリーと同じものである。ここでは、アークに音声片を割り当てたが、ノードに割り当てることもできる。語彙の終端にあたるノードには、その語彙の終端であることがわかるようにしておく。このようなノードをリーフノードと定義する。図3ではリーフノードを黒丸で表している。また、ツリーの深さを、根から数えて第1段、第2段、…と数えるとする。

【0056】音素マージ音声片ツリー11は、基本音声片ツリーのうち第1段～第n段までの音素をマージすることにより、語頭のツリーの広がりを小さくしたものである。第n+1段以降は基本音声片ツリーそのままである。

【0057】第1段～第n段までの音素マージは以下の

方法で行う。日本語の母音は、5種類しかなくこれらを識別することは比較的容易であるが、子音はカテゴリ数も多く識別が難しい。そこで、子音は音素群（無声破裂音、摩擦音、有性破裂音など）毎にまとめてマージし、同じ音素群内の子音は区別をしないとする。すなわち、子音は、音素の区別は行わず、無声破裂音や摩擦音のような音素群の区別しか行わない。語頭音素が1文字違うだけの「きりゅう」と「ちりゅう」は区別を付けずに照合することになる。

【0058】音響特徴の似通った子音の音素群内でのマージを行うため、マージによる誤差が少なく、しかも異なる音素群間の識別は音響特徴がかけ離れているため容易である。したがって、正解候補が枝刈られることはほとんどなく認識性能の低下が小さい。本実施の形態では、子音を図5のような4つのカテゴリに分ける。

【0059】音素をマージするとことにより音声片もマージされる。CVは後続母音が同じ場合に、VCは先行音素が同じ場合にマージする。音素群毎に音素をマージして得られる音声片を音素マージ音声片と定義する。音素マージ音声片のマージ方法と表記法の例を図6に示す。

【0060】基本音声片ツリーのうち、第1段～第n段までの音声片を、音素マージ音声片とすることにより、同じ音素マージ音声片を割り当てられたアーク同士をマージして語頭付近の広がりの小さいツリーにすることができる。これが音素マージ音声片ツリーである。

【0061】図3の基本音声片ツリーを、第1段～第3段（n=3）までの音声片をマージして音素マージ音声片ツリーに変換したのが図7である。図7の音素マージ音声片ツリーは、図3の音声片ツリーに比べ、語頭付近のツリーの広がりが狭くなっている。nを1とすると語頭の1番目の音声片だけをマージしたことになり、∞とするとすべての音声片をマージすることになる。nの大きさは、計算量がリアルタイムで収まる程度に決めておく効率が良い。音素マージ音声片ツリーでは、「きりゅう」と「ちりゅう」のようにリーフノードに複数の語彙が割り当てられることがある。

【0062】次に、本発明の実施の形態1における音声認識装置について、図2のフローチャートを参照しながらその動作を説明する。

【0063】図2において、音声片標準パターン15は、あらかじめ多数話者が発声した学習データから学習し、音声片毎に求めておく。また、音素マージ音声片標準パターン16は、マージする音声片すべての学習データから学習することにより求められる。たとえば、音声片/{p, t, k, c} i/に対する標準パターンは、/p i/（ピ）、/t i/（ティ）、/k i/（キ）、/c i/（チ）のすべての学習データから学習することにより得られる。これを、あらかじめすべての音素マージ音声片について求めておくものとする。

【0064】本実施の形態では、特徴パラメータベクトルの出現確率が複数のガウス分布の和（これを混合分布と呼ぶ）で近似できると仮定し、学習データから、標準パターンのフレームごとにガウス分布の平均値ベクトルおよび共分散行列を求め、これを標準パターンとする。

【0065】音素マージ音声片ツリー11は、あらかじめ認識対象語彙セット10から、ツリー展開処理S07において作成しておく。

【0066】まず、音響分析処理S01は、入力された未知音声信号を分析時間（以下フレームと呼ぶ）毎にD個の特徴パラメータに変換される。特徴パラメータとしては、線形予測分析によるLPCケプストラム係数、LPCメルケプストラム係数、メル線形予測分析によるメルLPCケプストラム係数、メルスケールフィルタバンクによるメル周波数ケプストラム係数（MFCC）など、音声認識に適したものならばどのようなものを用いても良い。

【0067】照合処理S02では、音素マージ音声片ツリー11にしたがって音素マージ音声片標準パターン16および音声片標準パターン15を接続しながら、音響分析処理S01からの未知入力音声の特徴パラメータ時系列と標準パターンとの照合を行う。照合は、入力フレーム同期のビームサーチを用いたDPマッチングにより行う。DPマッチングの方法およびビームサーチの方法は、従来例と同じであるため説明を省略する。この照合を一回目の照合と呼ぶ。

【0068】なお、本実施の形態では、音素マージ音声片ツリー11はあらかじめ作成しておくとしたが、ビームサーチDPを行いながら動的にツリー展開してもよい。

【0069】入力フレーム同期のビームサーチを用いたDPについて、その概念図を表したものが図9である。図9において、横軸は入力音声のフレーム、縦軸は音素マージ音声片ツリーにしたがって接続した音声片標準パターンのフレームを表している。辞書である縦軸はツリー状になっている。入力音声とツリー状の辞書のDPマッチングは、図9のようなツリー状のDP面上での入力と標準パターンの最適な経路を求めながらスコアを算出するものである。このツリー状のDP面は、第1段～第n段までが音素マージされており枝の広がり小さくなっている。

【0070】DPマッチングは、ビームサーチにより入力フレーム同期にDP経路の枝刈りを行う。ビーム内に残る格子点候補数はDP面のすべての格子点数に比べはるかに少ないため、このDP面は実際にメモリ上に持つ必要はなく、仮想的なものである。

【0071】発声開始からしばらくすると、発声内容と似ていない辞書のDPパスの累積スコアは、正解パスの累積スコアに比べ十分小さな値になり枝刈られるため、格子点候補数は急激に減少する。したがって、それまで

の間の格子点候補数を抑えることが全体の計算量削減につながる。第一の実施の形態のように語頭付近のツリーの広がりを抑えることにより、発声開始付近のビーム内に残る格子点候補数は大幅に削減することができる。

【0072】判定処理S03では、DPマッチングにより最も累積スコアの高かったリーフノード（最大ゆう度リーフノード）を求め、これに対応する語彙が一意に決まるかどうかの判定を行う。もし、一意に決まる場合（Y）、すなわち最大ゆう度リーフノードに対応する語彙が1個しかない場合は、その語彙を認識結果として出力する。もし、一意に決まらない場合（N）、すなわち最大ゆう度リーフノードに対応する語彙が複数存在する場合には、次のような方法で認識結果を決定する。

【0073】再照合候補抽出処理S05において、再照合候補を抽出する。本実施の形態では、再照合候補を最大ゆう度リーフノードに対応する語彙とする。他の方法としては、最大ゆう度リーフノードだけでなく、ビーム内に残った累積スコアの上位K個のリーフノードに対応する語彙をすべて再照合候補とする方法もある。

【0074】次に、再照合用ツリー展開処理S06において、再照合候補の語彙に対して音素マージを行わない第1段～第n段までの音声片ツリーを展開する。この音声片ツリーを再照合用前半音声片ツリー13とする。再照合用前半音声片ツリー13は、第1段～第n段で認識語彙が一意に決まる。そこで、第n段の終端ノードにその語彙を割り当てておく。再照合候補が「めぐろ」「めむろ」「ねむろ」の3単語、 $n=3$ であった場合の、再照合用前半音声片ツリー13の例を図8に示す。

【0075】本実施の形態では、照合処理S02をあらかじめDPマッチングを行う際、第1段の始端ノードに対応する入力フレーム位置Fsと、第n段の終端ノードに対応する入力フレーム位置Feを記憶しておく必要がある。

【0076】前半再照合処理S04では、再照合用前半音声片ツリー13にしたがって接続した音声片標準パターン15と、フレームFsからフレームFeまでの入力音声とを、DPマッチングにより再照合する。再照合の場合は、認識対象語彙が少ないためビームサーチは必ずしも行わなくてもよい。再照合の結果、最も累積スコアの高かった再照合用ツリーの第n段の終端ノードに対応する語彙を認識結果として出力する。

【0077】なお、ビーム内に残った累積スコアの上位K個のリーフノードに対応する語彙をすべて再照合候補とする方法の場合には、発声の前半部分のスコア、すなわち再照合の結果求める入力フレームFsからフレームFeまでのスコアSaと、発声の後半部分のスコア、すなわち一回目の照合の結果求める入力フレームFe+1から発声の終端フレームまでのスコアSbとの和Sを、再照合候補の語彙すべてについて求め、Sの最も大きい語彙を認識結果とする。

10

20

30

40

50

【0078】本実施の形態では、音素マージを行うのは一律第1段～第n段としたが、すべての段においても良い。また、ツリーの密集しているところは深くしたりするなど部分的に変えても良い。再照合を行うのも一律第n段の終端ノードまでではなく、単語が一意に決まるノードまでとしてもよい。すべての段において音素マージを行う場合には一回目の照合では音素マージしていない音声片標準パターン15を使用する必要はない。

【0079】また、本実施の形態では、最大ゆう度リーフノードに対応する語彙が1個であった場合は再照合を行わないとしたが、その場合でも、最大ゆう度リーフノードだけでなく、ビーム内に残った累積スコアの上位K個のリーフノードに対応する語彙をすべて再照合候補としてもよい。

【0080】以上のように、本実施の形態によれば、第一段から第n段までの音声片について、同じ音素群に属する子音をマージした音素マージ音声片ツリーを用いることにより、一回目の照合における計算量を大幅に削減することができ、再照合を行ったとしても全体の計算量は大幅に削減することができるという効果があります。

【0081】また、この方法では似た音素を区別しないで認識するため、一回目の照合で正解候補が漏れる可能性が低く、認識性能を劣化させずに計算量を削減することができるという効果があります。

【0082】さらに、本実施の形態では、再照合は1回目の照合で第1段の始端ノードに対応する入力フレーム位置Fsと第n段の終端ノードに対応する入力フレーム位置Feを記憶しておき、FsからFeまでの間でのみ再照合を行えばよいので、再照合にかかる計算量は非常に少なくすむという効果があります。

【0083】（実施の形態2）次に、本発明の実施の形態2の音声認識装置について、図10のフローチャートを参照しながらその動作を説明する。

【0084】実施の形態1と異なるのは、再照合用前半音声片ツリー13が再照合用音声片ツリー14に、前半再照合処理S04が再照合処理S21になっていることである。再照合用音声片ツリー14は、実施の形態1と異なり、第1段～第n段だけではなく、単語終端までを表すツリーになっている。

【0085】実施の形態2の動作は、ほぼ実施の形態1と同じであるため、異なる部分についてのみ説明する。

【0086】実施の形態1では、再照合は、1回目の照合のときに音素マージ音声片ツリーの音素マージを行った第1段～第n段に対応していた入力区間についてのみ行ったが、実施の形態2では、発声区間全体について再照合を行う。

【0087】再照合用ツリー展開処理S09では、再照合候補の語彙に対して音素マージを行わない音声片ツリーを展開する。この音声片ツリーを再照合用音声片ツリー14とする。再照合用音声片ツリー14は、第1段～

第n段までではなく、単語終端までを表す音声片ツリーである。

【0088】再照合候補が「めぐろ」「めむろ」「ねむろ」の3単語であった場合の、再照合用音声片ツリーの例を図11に示す。

【0089】本実施の形態では、照合処理S02で、第1段の始端ノードに対応する入力フレーム位置および、第n段の終端ノードに対応する入力フレーム位置を記憶しておく必要はない。

【0090】再照合処理S21では、再照合用音声片ツリー14にしたがって接続した音声片標準パターン15と、入力音声の発声開始から発声終了までを、DPマッチングにより再照合する。再照合の場合は、実施の形態1と同様、認識対象語彙が少ないためビームサーチは必ずしも行わなくてよい。

【0091】再照合処理S21の結果、最も累積スコアの高かった再照合用ツリーのリーフノードに対応する語彙を認識結果として出力する。

【0092】以上のように、実施の形態2によれば、1回目の照合で第n段の終端ノードに対応する入力フレーム位置が最適な位置ではなかった場合に、発声区間の開始から終了までを再照合することにより、より精密な照合を行うことができるため、実施の形態1に比べさらに認識性能が向上するという効果があります。

【0093】また、実施の形態2では、第1段の始端ノードに対応する入力フレーム位置Fsと第n段の終端ノードに対応する入力フレーム位置Feを記憶しておく必要がないため1回目の認識処理およびメモリ容量は実施の形態1に比べ少なくすむという効果があります。

【0094】また、実施の形態2のように、発声区間の開始から終了までを再照合する場合には、再照合の距離尺度は1回目のものとまったく違うものを用いてもかまわない。そのため、再照合の際には数単語のみより精密に認識できる方法を用いて、より高い認識性能を得ることもできる。

【0095】（実施の形態3）次に、本発明の実施の形態3における音声認識装置について、図12のフローチャートを参照しながらその動作を説明する。

【0096】実施の形態1と異なるのは、音素マージ音声片ツリー11がラフ音声片ツリー12に、音素マージ音声片標準パターン16がラフ音声片標準パターン17に、音声片標準パターン15が精密音声片標準パターン18になっていること、および判定処理S03が不要なことである。

【0097】精密音声片標準パターン18は、実施の形態1の音声片標準パターン15と同じものである。実施の形態3では、ラフ音声片と対比づけるために通常の音声片を精密音声片と呼ぶことにする。

【0098】ラフ音声片ツリー12およびラフ音声片標準パターン17について以下に説明する。ラフ音声片

10

20

30

40

50

は、音声片の標準パターンの精度を粗くしたものと定義する。その方法としては、次の二つが考えられる。

【0099】一つ目は、ラフ音声片1つの音声片あたりにかかる距離計算量を精密音声片1つあたりにかかる計算量に比べ削減する方法である。具体的には、ラフ音声片標準パターンの、フレーム数を少なくする方法、ガウス分布の混合数を削減する方法、ガウス分布の共分散行列を共通化して共分散行列の種類数を削減する方法などが考えられる。この方法では音声片ツリーの形状は変わらない。

【0100】二つ目は、認識結果が一意に決まる範囲内で、異なる音韻環境の音声片をマージする方法である。この方法によっても、ツリーのアークとノードが減るため計算量を削減することができる。たとえば、VCは母音部分が同じであれば後続子音が異なっても1つの音声片にマージするなどが考えられる。この方法では、音声片ツリーの形状が変わることがある。音声片の単位として音素を用いる場合は、音素の前後の音素環境によって異なる音声片とすることが多いが、中心音素が同じ場合には1つの音声片にマージすることにより、ツリーの広がりを大幅に抑えることが可能になる。当然のことながら中心音素が同じであればマージを行っても認識結果は必然的に一意に決まる。

【0101】実施の形態3では、標準パターンのフレーム数を削減する方法と、母音部分が同じで後続子音の異なるVCをマージする方法の両方を行う。前者は音声片記号の上にバーをつけて表記し、後者は子音部分をアスタリスクで置き換えて表記することとする。

【0102】図13は、図3の基本音声片ツリーを、第1段〜第3段($n=3$)までの音声片をラフ音声片としたラフ音声片ツリーである。第4段以降は、基本音声片ツリーと同じである。ツリーの形状は図2と若干変わっている。なお、VCのマージは、後続子音が同じ音素群の場合のみに限っても良い。

【0103】ラフ音声片標準パターン17は、以下のようにならかじめ学習し求めておく。標準パターンのフレーム数をもとのフレーム数の半分に減らして学習する。さらにVCは母音部分が同じ音声片すべての学習データから学習する。たとえば、音声片/e*/に対する標準パターンは、母音部分が/e/で後続子音が異なる音声片/em/, /en/, /eg/, /eb/, …のすべての学習データから学習することにより得られる。

【0104】実施の形態3の動作は、ほぼ実施の形態1と同じであるため、異なる部分についてのみ説明する。

【0105】照合処理S02では、ラフ音声片ツリー12にしたがってラフ音声片標準パターン17および精密音声片標準パターン18を接続しながら、実施の形態1と同様にして、未知入力音声の特徴パラメータ時系列と標準パターンとの照合を行う。

【0106】照合を行った後、再照合候補抽出処理S0

5で、再照合候補を抽出する。本実施の形態では、ビーム内に残った累積スコアの上位K個のリーフノードに対応するK個の語彙を再照合候補とする。実施の形態1と同様に再照合候補に対して再照合用前半音声片ツリー13を展開し、前半再照合処理S04において発声前半部分について精密な音声片標準パターンで照合を行う。

【0107】再照合の結果求める発声の前半部分のスコア S_a と、一回目の照合の結果求める発声の後半部分のスコア S_b との和 S を、再照合候補の語彙すべてについて求め、 S の最も大きい語彙を認識結果とする。

【0108】本実施の形態では、音素マージを行うのは一律第1段〜第 n 段としたが、ツリーの密集しているところは深くしたりするなど部分的に変えても良い。再照合を行うのも一律第 n 段の終端ノードまでではなく、単語が一意に決まるノードまでとしてもよい。

【0109】以上のように、実施の形態3によれば、音声片の標準パターンの精度を粗くしたラフ音声片ツリーを用いることにより、ラフ音声片の照合にかかる計算量が少なくてすむため、一回目の照合における計算量を大幅に削減することができ、再照合を行っても全体の計算量は削減できる。

【0110】発声開始直後の計算量の多いところは粗い照合を、発声開始後しばらくしてから計算量の少ないところは精密な照合をするため効率が良いという効果があります。

【0111】(実施の形態4)次に、本発明の実施の形態4の音声認識装置について、図14のフローチャートを参照しながらその動作を説明する。

【0112】実施の形態3と異なるのは、前半再照合処理S04が再照合処理S21に、再照合用前半音声片ツリー13が再照合用音声片ツリー14になっていることである。実施の形態4は、実施の形態3と実施の形態2の組み合わせである。再照合処理S21と再照合用音声片ツリー14は、実施の形態2と同じである。

【0113】実施の形態4の動作は、ほぼ実施の形態3と同じであるが、再照合用ツリー展開処理S06において再照合用音声片ツリー14を作成し、音声の前半部分のみ前半再照合処理S21において再照合を行うところは、実施の形態2と同じである。

【0114】(実施の形態5)次に、本発明の実施の形態5における音声認識装置について、図15フローチャートを参照しながらその動作を説明する。

【0115】実施の形態3と異なるのは、前半再照合処理S04、再照合候補抽出処理S06、再照合用ツリー展開処理S06、再照合用前半音声片ツリー13が不要なことである。

【0116】実施の形態3では、一回目の照合で認識結果は一意に決まるため、再照合を行わなくても認識結果を出力することができる。そこで再照合を行わずに認識結果を出力するとしたのが実施の形態5である。実施の

形態 5 の動作は、再照合を行わずに一回目の照合の結果をそのまま認識結果とする以外は実施の形態 3 と同じである。

【0117】実施の形態 5 では、実施の形態 3 に比べ、再照合を行わないため認識性能は劣化するが計算量は大幅に削減できる。その場合でも音声片の標準パターンの精度を粗くするのは探索空間の広い語頭付近だけであるため、一律に音声片の標準パターンの精度を粗くするよりも効率的な計算量削減が図れる。また、実施の形態 5 では、再照合の必要がないため、入力音声の特徴パラメータ情報を記憶しておく必要がなく容量も小さくてすむという利点がある。

【0118】

【発明の効果】以上のように本発明は、特徴の似ている音素をマージした音声片を用いて照合を行い、認識結果が一意に決まらなかった場合にのみ再照合を行うことにより、認識性能を落とさずに計算量を削減することができる。

【0119】また、語頭付近について音声片の標準パターンの精度を粗くしたラフ音声片ツリーを用いて照合を行ったのち、精密な音声片標準パターンを用いて再照合することによって認識性能を落とさずに効率よく計算量を削減することができる。

【0120】さらに、音声片の標準パターンの精度を粗くしたラフ音声片ツリーを用いて照合を行い、再照合を行わない場合には、認識性能の劣化を最小限に抑え計算量を大幅に削減することができる。

【0121】さらに、1 回目の照合で第 1 段の始端ノードに対応する入力フレーム位置 F_s と第 n 段の終端ノードに対応する入力フレーム位置 F_e を記憶しておき、 F_s から F_e までの間でのみ再照合を行う場合には、再照合にかかる計算量を抑えることができる。

【0122】また、発声区間の開始から終了までを再照合する場合には、より精密な再照合が行えるため認識性能の劣化が少なくすみ、1 回目の照合方法とまったく違うものを用いてもかまわないため、より精密な手法で再照合を行った場合にはより高い認識性能を得ることもできる。

【図面の簡単な説明】

【図 1】本発明の実施の形態におけるコンピュータを用いた音声認識装置の構成図

【図 2】本発明の実施の形態 1 における音声認識装置のフローチャート

【図 3】本発明の実施の形態 1 における基本音声片ツリーを示す図

【図 4】本発明の実施の形態 1 における音声片列を示す図

【図 5】本発明の実施の形態 1 における音素群の定義を説明する図

【図 6】本発明の実施の形態 1 における音声片のマージを説明する図

【図 7】本発明の実施の形態 1 における音素マージ音声片ツリーを示す図

【図 8】本発明の実施の形態 1 における再照合用前半音声片ツリーを示す図

【図 9】本発明の実施の形態 1 における仮想 DP 面を説明する図

【図 10】本発明の実施の形態 2 における音声認識装置のフローチャート

【図 11】本発明の実施の形態 2 における再照合用音声片ツリーを示す図

【図 12】本発明の実施の形態 3 における音声認識装置のフローチャート

【図 13】本発明の実施の形態 3 におけるラフ音声片ツリーを示す図

【図 14】本発明の実施の形態 4 における音声認識装置のフローチャート

【図 15】本発明の実施の形態 5 における音声認識装置のフローチャート

【図 16】従来の音声認識装置の構成図

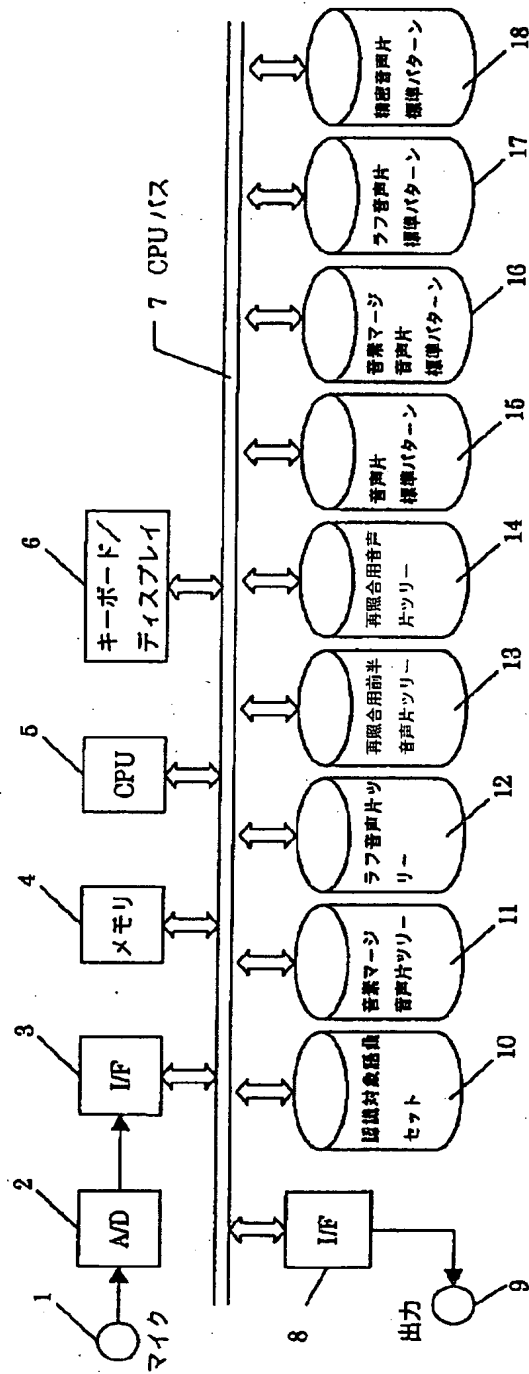
【図 17】従来例における音声認識装置のフローチャート

【図 18】従来例の計算量を説明する図

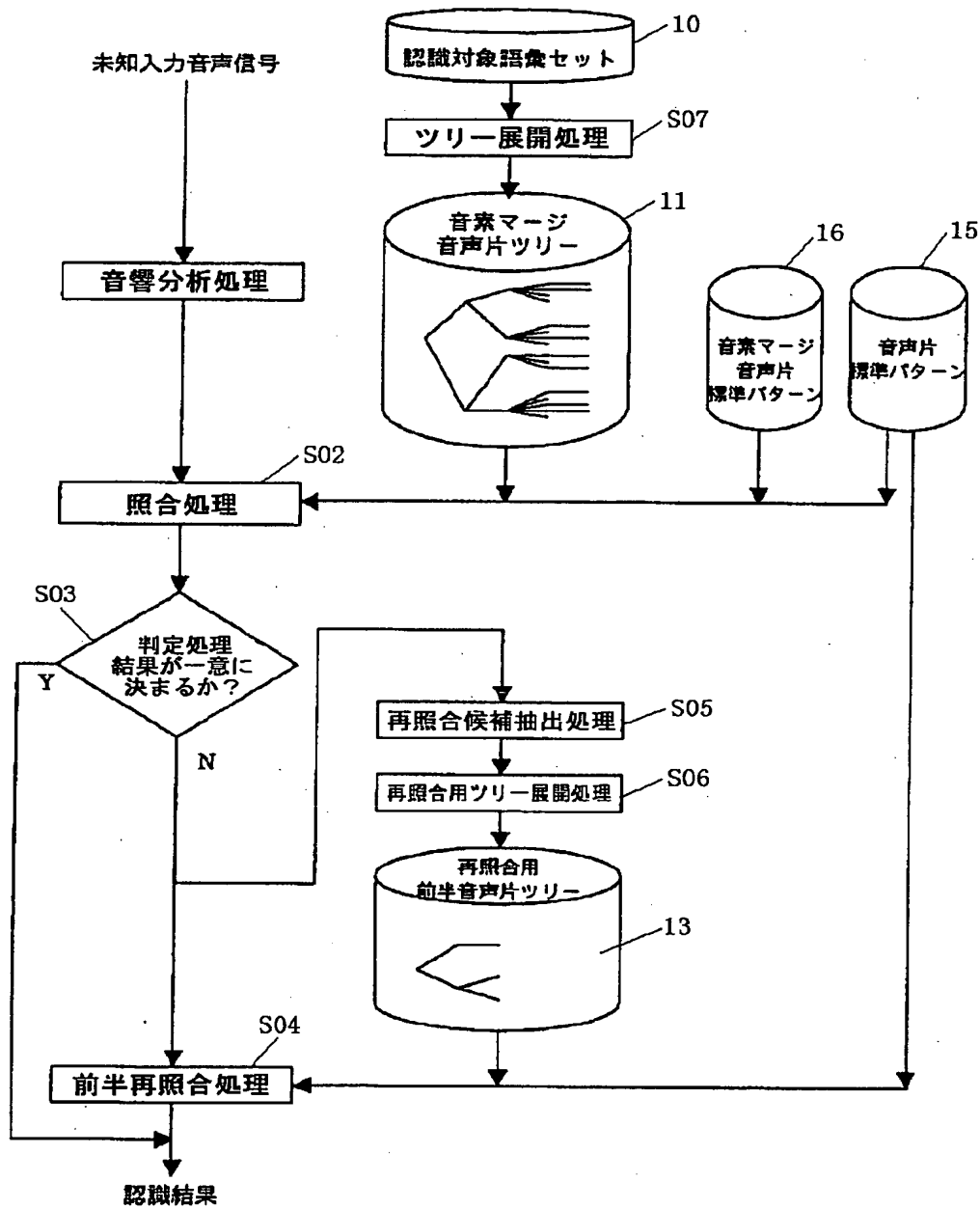
【符号の説明】

- 1 マイク
- 2 A/D
- 3 インタフェース (I/F)
- 4 メモリ
- 5 CPU
- 6 キーボード/ディスプレイ
- 7 CPUバス
- 8 I/F
- 9 出力
- 10 認識対象語彙セット
- 11 音素マージ音声片ツリー
- 12 ラフ音声片ツリー
- 13 再照合用前半音声片ツリー
- 14 再照合用音声片ツリー
- 15 音声片標準パターン
- 16 音素マージ音声片標準パターン
- 17 ラフ音声片標準パターン
- 18 精密音声片標準パターン
- 19 音声片ツリー

【図1】



【図2】



【図5】

無声破裂音・無声破擦音群

{/p/, /t/, /k/, /c/}

無声摩擦音群

{/s/, /h/, /z/}

鼻音群

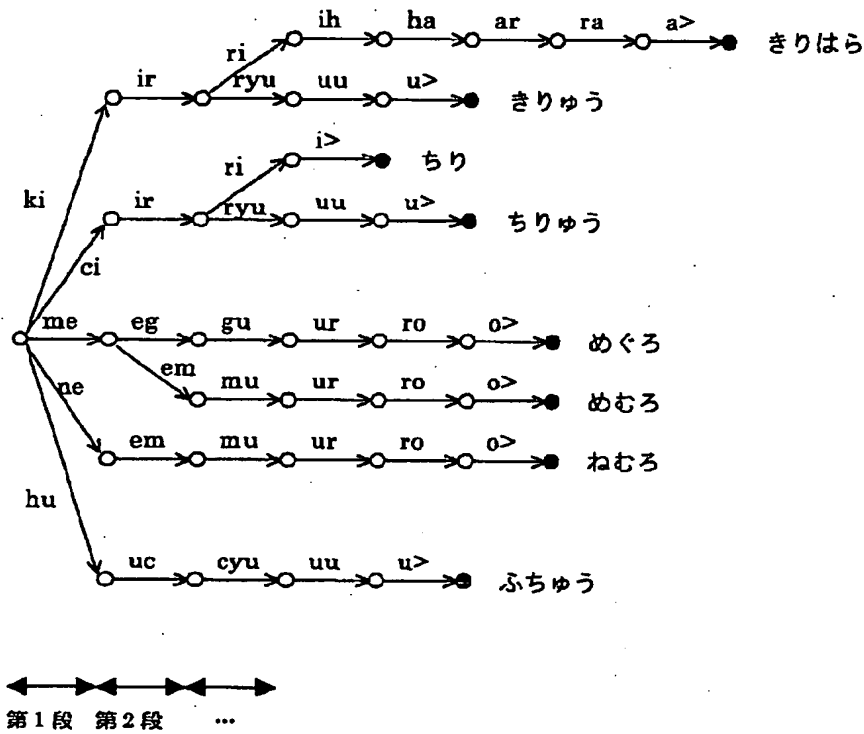
{/m/, /n/}

有性破裂音・流音群

{/b/, /d/, /g/, /r/}

【図3】

基本音声片ツリー



【図4】

音声片列

ki - ir - ri - ih - ha - ra - a>
 ki - ir - ryu - uu - u>
 ci - ir - ri - i>
 ci - ir - ryu - uu - u>
 me - eg - gu - ur - ro - o>
 me - em - mu - ur - ro - o>
 ne - em - mu - ur - ro - o>
 hu - uc - cyu - uu - u>

ただし">"は語尾記号

【図6】

CVのマージ例:

もとの音声片表記
 /pi/ , /ti/ , /ki/ , /ci/

マージ音声片表記
 → /{p,t,k,c}i/

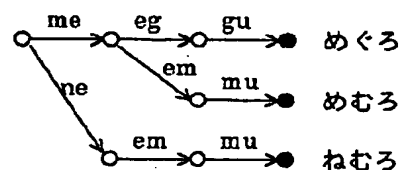
VCのマージ例:

もとの音声片表記
 /a/ , /at/ , /ak/ , /ac/

マージ音声片表記
 → /a{p,t,k,c}/

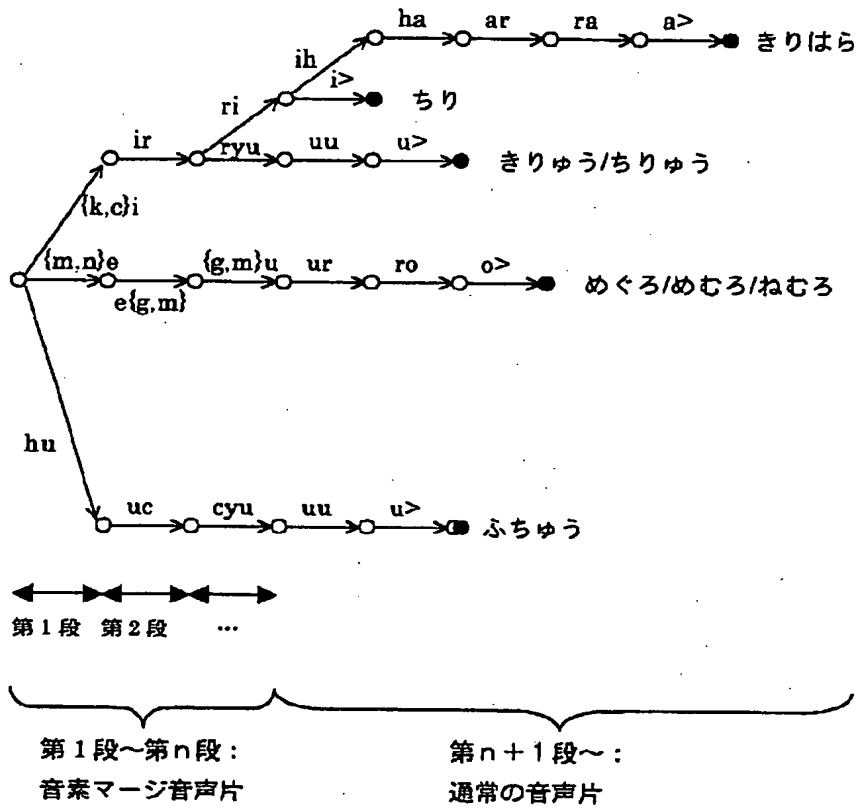
【図8】

再照合用前半音声片ツリー

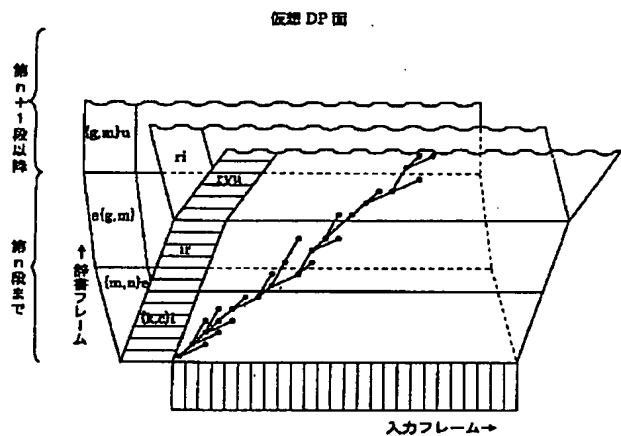


【図7】

音素マージ音声片ツリー

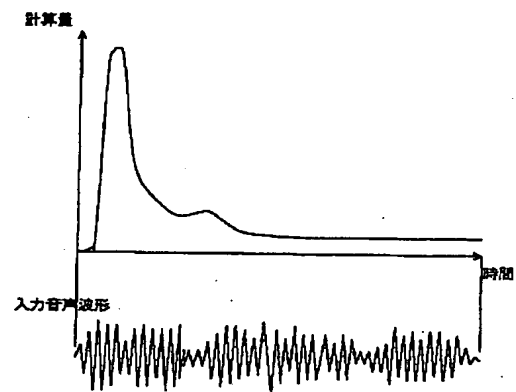


【図9】

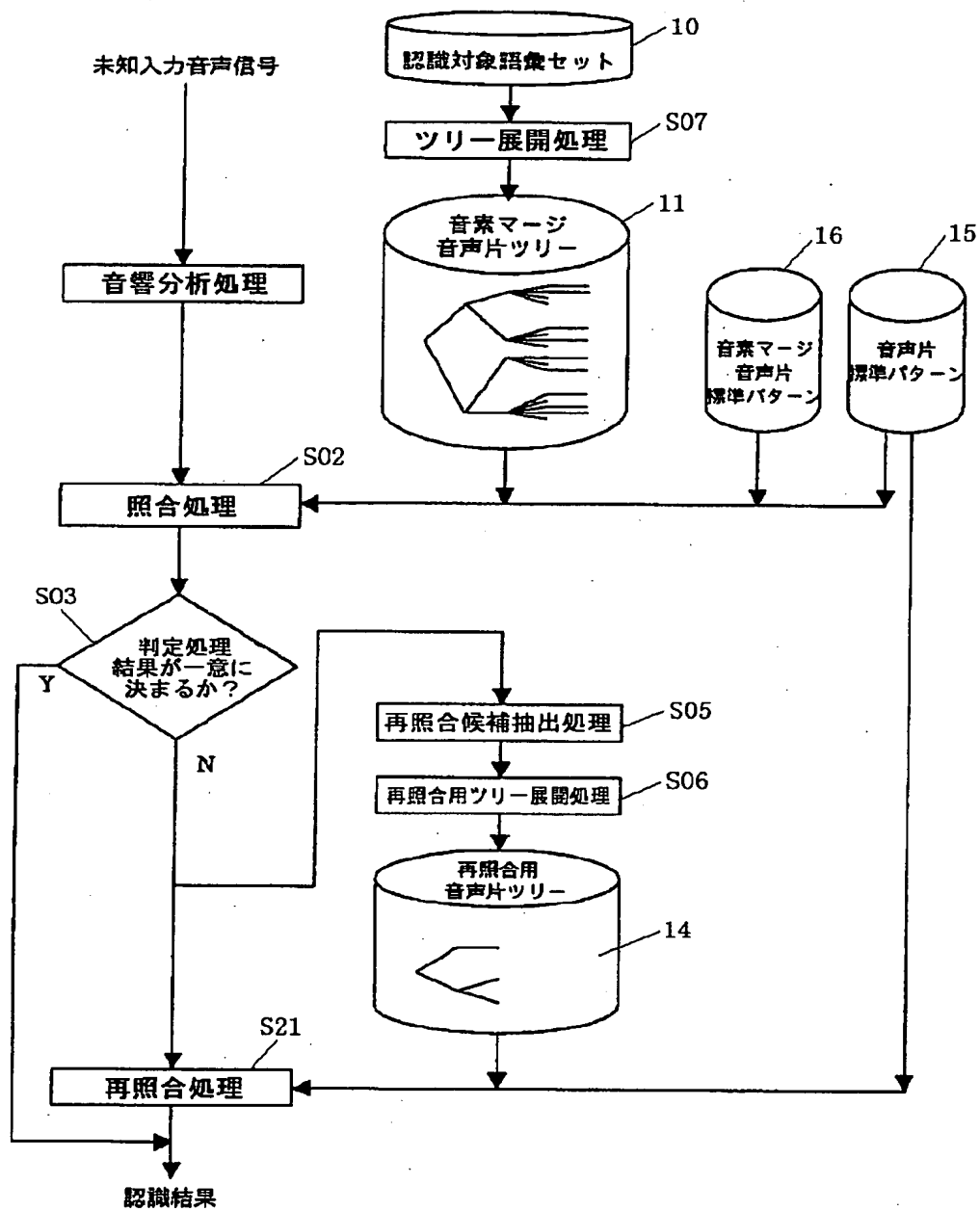


● = ビーム内に残った格子点候補

【図18】

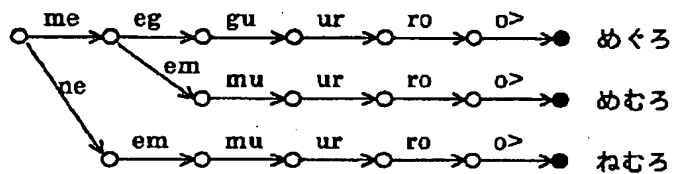


【図10】

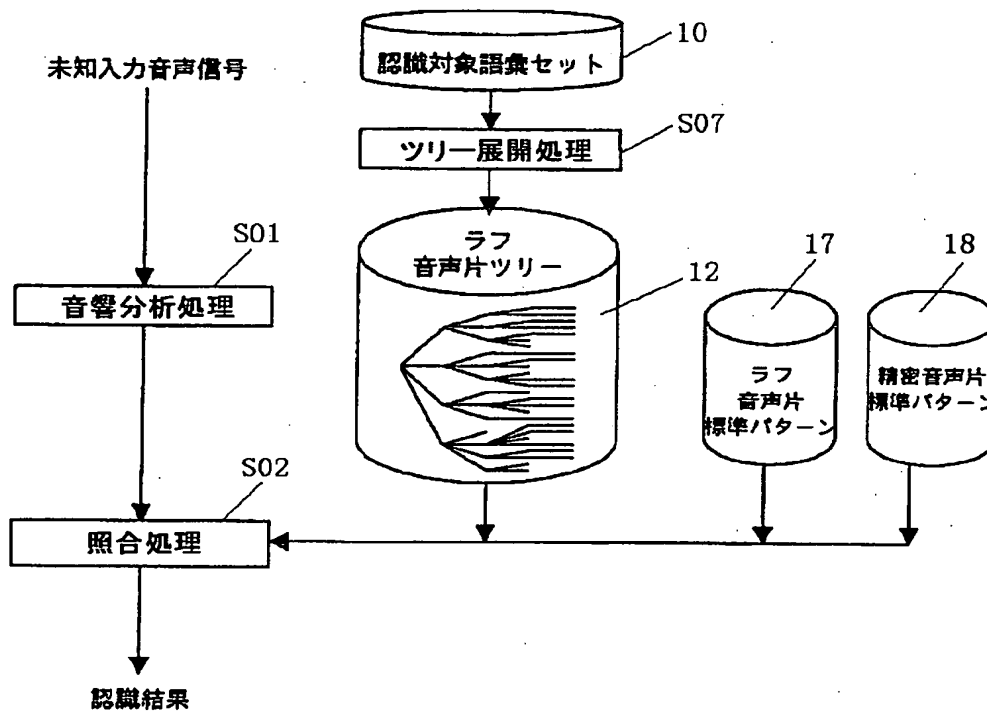


【図11】

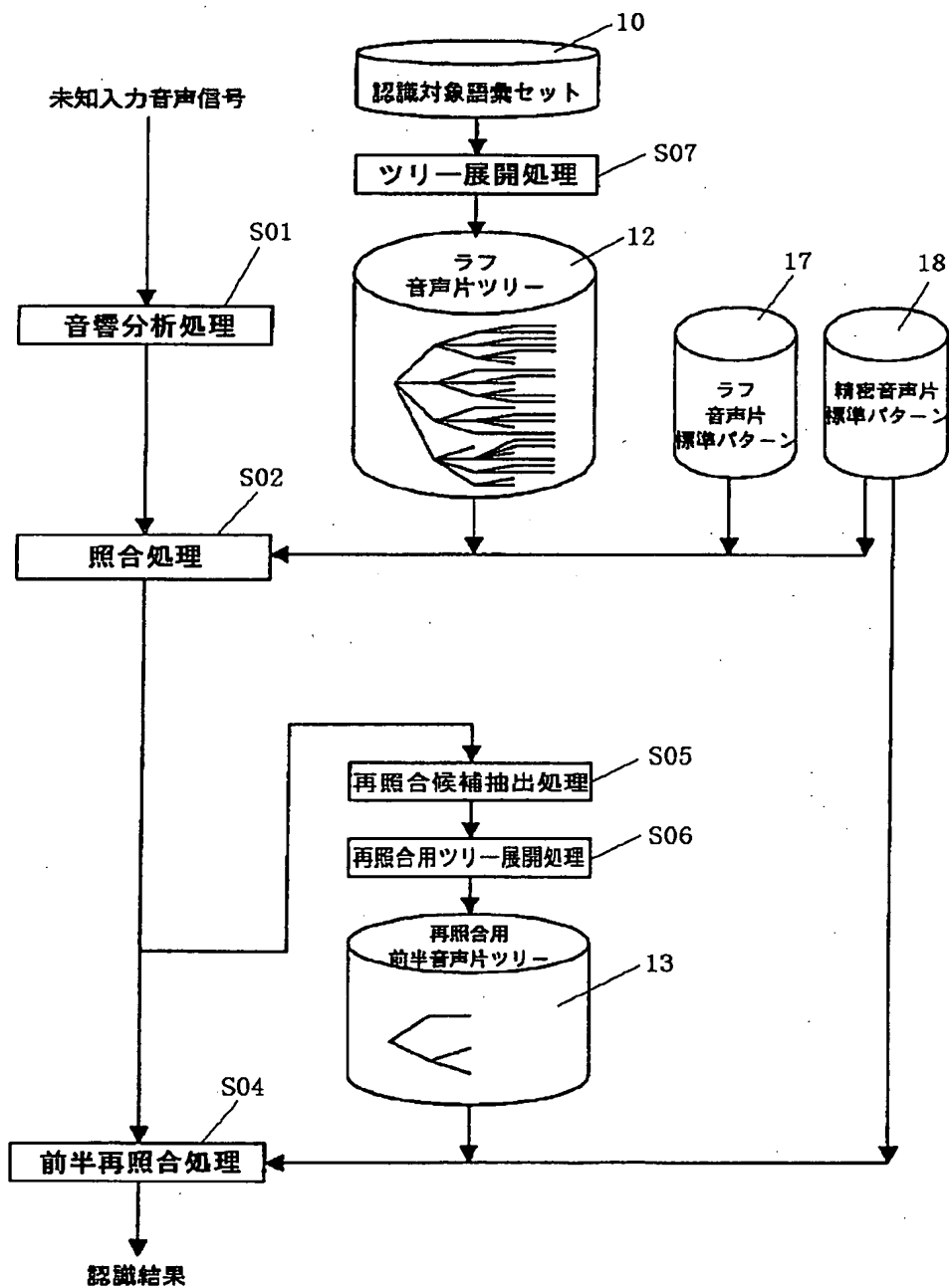
再照合用音声片ツリー



【図15】

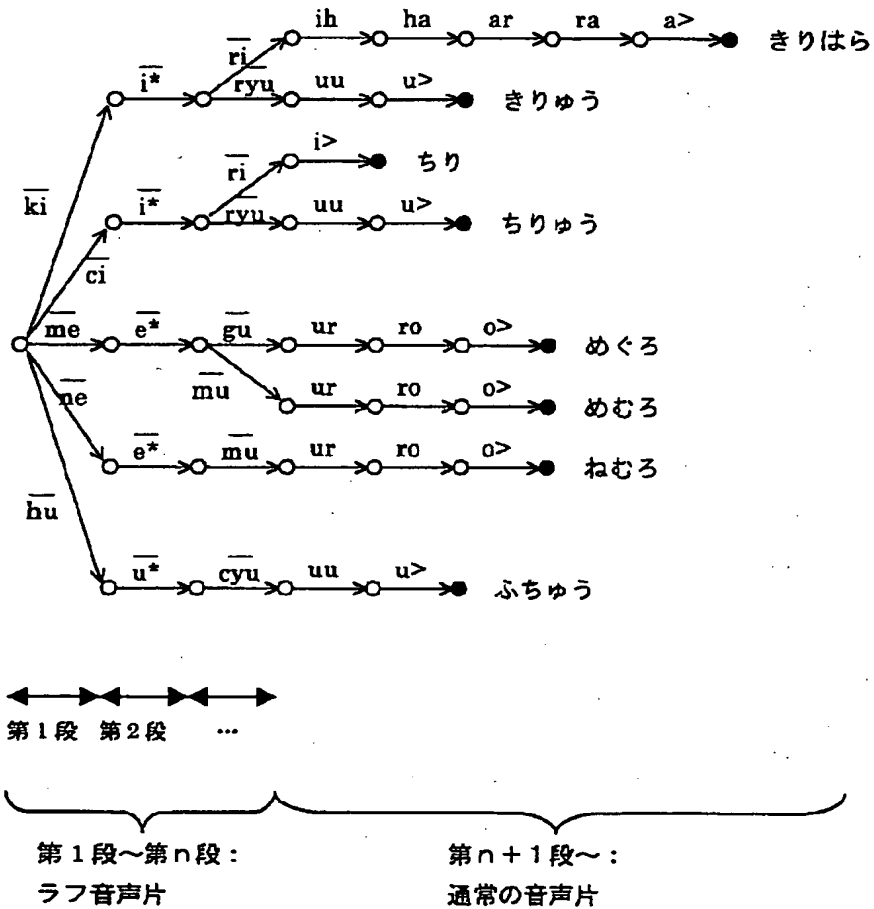


【図12】



【図13】

(a) ラフ音声片ツリー



(b)

ラフ音声片の表記

(1) 標準パターンのフレーム数を削減する場合:

音声片記号の上にバーをつけて表記

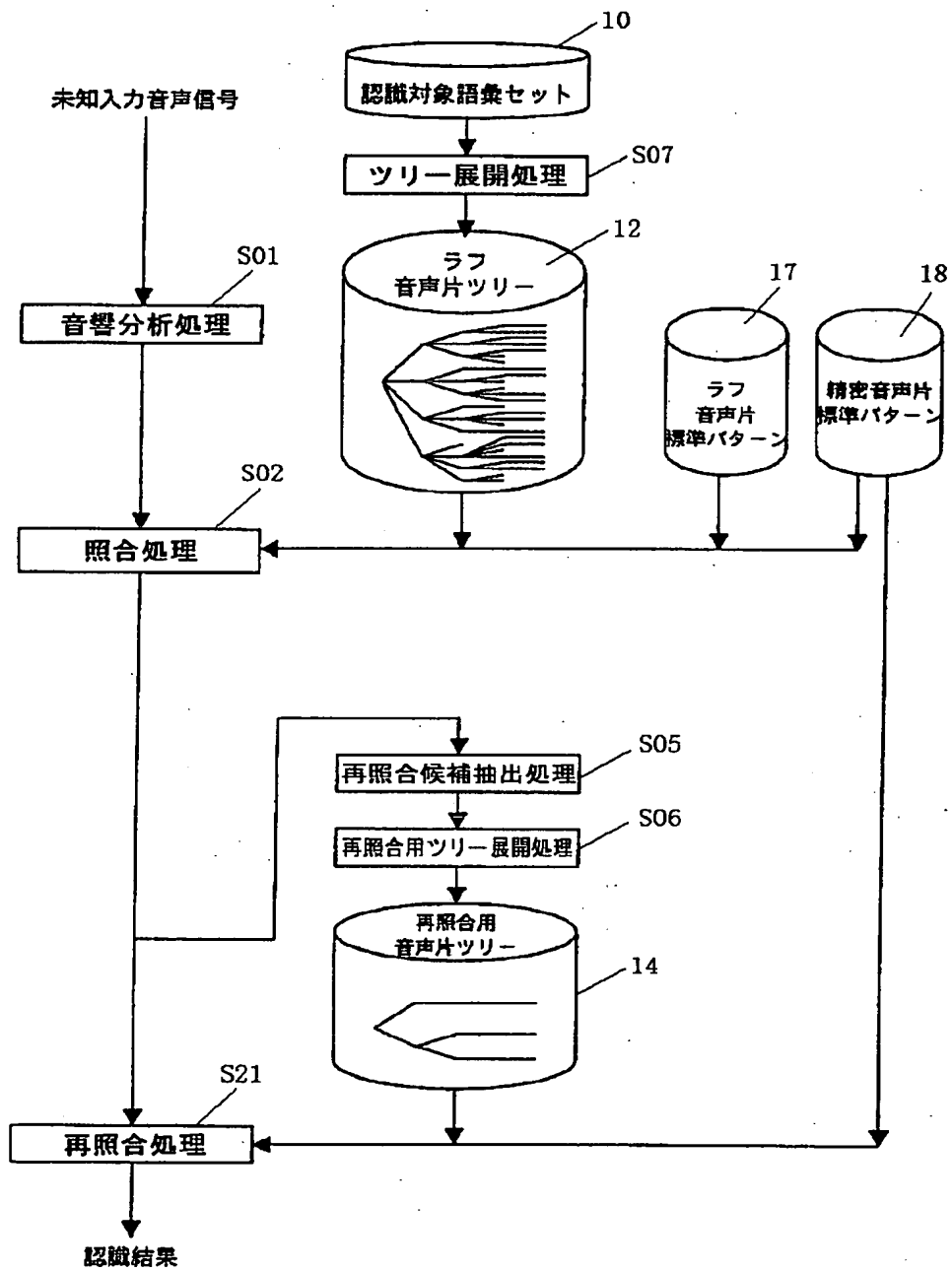
(例) ki → $\overline{\text{ki}}$

(2) 母音部分が同じで後続子音の異なる VC をマージする。

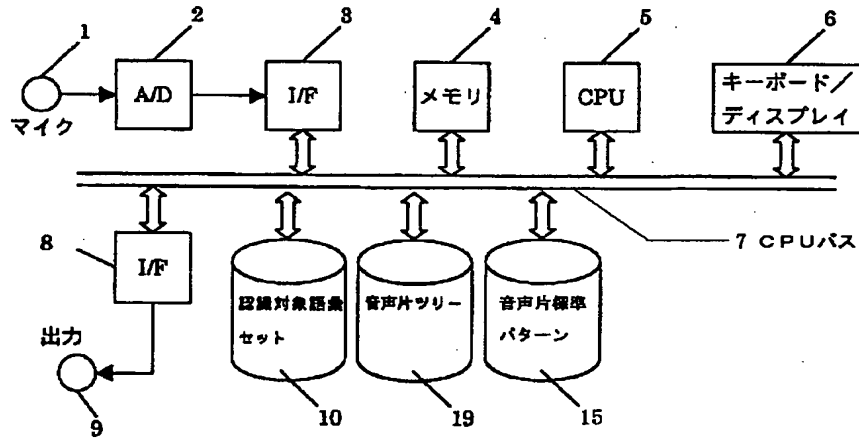
子音部分をアスタリスクに置き換えて表記

(例) em, en, eg, eb, ... → e^*

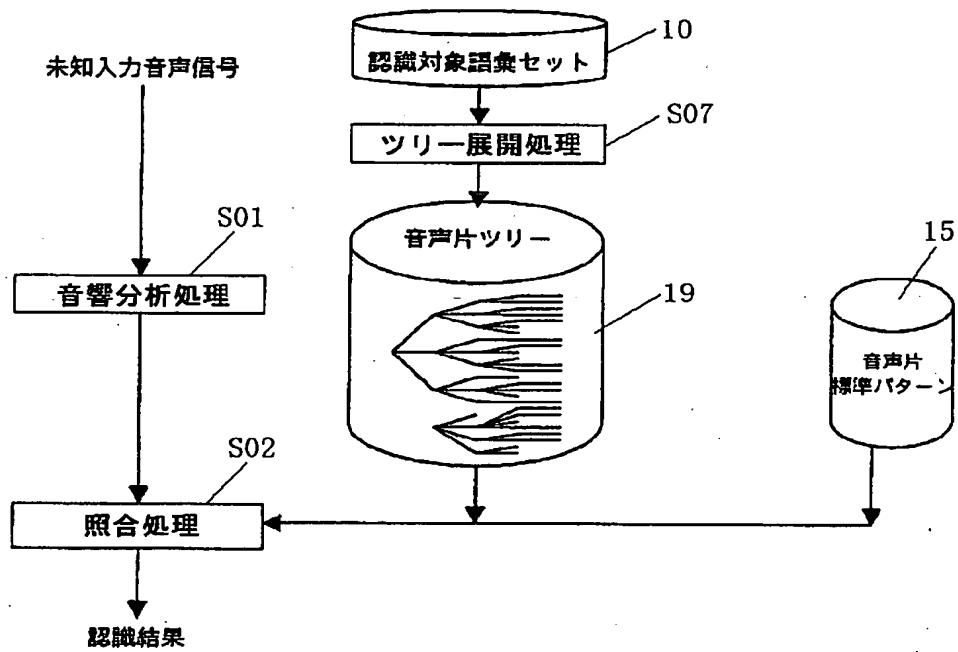
【図14】



【図16】



【図17】



[0050] (Embodiment 1)

Fig. 1 is a block diagram of a speech recognition apparatus of Embodiment 1 of the present invention, which will hereinafter be described.

[0051] In Fig. 1, reference numeral 1 denotes a microphone for collecting speech, reference numeral 2 denotes an A/D converter, reference numeral 3 denotes an interface (I/F), reference numeral 4 denotes a memory, reference numeral 5 denotes a CPU, reference numeral 6 denotes a keyboard/display, reference numeral 7 denotes a CPU bus, reference numeral 8 denotes an I/F, reference numeral 9 denotes an output, reference numeral 10 denotes a recognition target dictionary set, reference numeral 11 denotes a phoneme merged speech segment tree, reference numeral 12 denotes a rough speech segment tree, reference numeral 13 denotes a first half speech segment re-collation tree, reference numeral 14 denotes a speech segment re-collation tree, reference numeral 15 denotes a speech segment standard pattern, reference numeral 16 denotes a phoneme merged speech segment standard pattern, reference numeral 17 denotes a rough speech segment standard pattern, and reference numeral 18 denotes an accurate speech segment standard pattern.

[0052] First, the phoneme merged speech segment tree 11

corresponding to a recognition dictionary in Embodiment 1 will be described with reference to Figs. 3, 4 and 5.

[0053] As a unit of a standard pattern, a phoneme segment, a phoneme, a syllable, CV/VC (a consonant + a vowel/ a vowel + a consonant), VCV, CVC and the like are conceivable. These minimum recognition units are referred to as speech segments. In the present embodiment, CV representing from the start of a consonant to the center of a vowel, VC representing from the center of a vowel to the end of the vowel, and VV representing from the center of a vowel to the center of a vowel serve as base units. Although VC includes only a vowel section, it is defined differently depending on the subsequent consonant.

[0054] For example, assume that the recognition target vocabulary includes 8 words: "kirihara", "kiryu", "chiri", "chiryu", "meguro", "memuro", "nemuro" and "fuchu", and when they are represented by speech segment sequences, the results are as shown in Fig. 4.

[0055] Fig. 3 is a diagram showing the above results in a simple tree structure. In the present embodiment, this is defined as a base speech segment tree. This is the same as the speech segment tree used in the prior art. Although speech segments are allocated to arcs here, they may also be allocated to nodes. Each node corresponding to an end of a vocabulary is made recognizable as the end of the

vocabulary. Such a node is defined as a leaf node. In Fig. 3, the leaf nodes are represented by solid circles. Further, the depth of the tree is to be counted from its root (e.g., a first step, a second step ...).

5 [0056] In the phoneme merged speech segment tree 11, phonemes from a first to nth steps of the base phoneme tree are merged, whereby expansion of initial phonemes of words in the tree is suppressed. The phoneme merged speech segment tree is the same as the base speech segment tree
10 subsequent from a (n+1)th step.

[0057] Phoneme merging from the first to nth steps is performed by the following method. There are only five vowels in Japanese so that it is relatively easy to distinguish them. On the other hand, it is very difficult
15 to distinguish consonants because the number of categories is large. Consonants are collectively merged for each phoneme group (e.g., unvoiced stop consonant, fricative consonant, voiced stop consonant groups and the like), and consonants within the same phoneme group are not
20 distinguished. That is, in each consonant, phonemes are not distinguished, and they are distinguished only by the phoneme group such as the unvoiced stop consonant or fricative consonant group. This means that "kiryu" and "chiryu", which are different by one letter in initial
25 phoneme, are to be collated without any distinction.

[0058] Since merging is performed between consonants having similar acoustic features in the same phoneme group, there is little error due to merging, and moreover, distinction between different phoneme groups is easy because their acoustic features are very different. Therefore, a correct candidate is hardly pruned so that deterioration of recognition performance is little. In the present embodiment, the consonants are divided into four categories as shown in Fig. 5.

[0059] By merging phonemes, speech segments are also merged. Merging occurs to CV when its subsequent vowel is the same, and merging occurs to VC when its antecedent phoneme is the same. A speech segment obtained by merging phonemes for each phoneme group is defined as a phoneme merged speech segment. Examples of the merging method of the phoneme merged speech segment and its notation are shown in Fig. 6.

[0060] Of the base speech segment tree, speech segments from the first to nth steps are converted to phoneme merged speech segments, whereby arcs to which the same phoneme merged speech segment is allocated are merged making it possible to obtain a tree with a reduced expansion in the vicinity of initial phonemes of words. This is the phoneme merged speech segment tree.

[0061] Fig. 7 is a phoneme merged speech segment tree

obtained by merging speech segments from the first to third steps ($n=3$) of the base speech segment tree of Fig. 3 and converting them to phoneme merged speech segments. In the phoneme merged speech segment tree of Fig. 7, expansion of initial phonemes of words in the tree is suppressed, compared with the speech segment tree of Fig. 3. Supposing that n is 1, first speech segments of the initial phonemes of the words are merged, and supposing that n is ∞ , all the speech segments are to be merged. It is efficient if the size of n is determined to an extent that the quantity of calculation is within the real-time processing. In the phoneme merged speech segment tree, a plurality of words may be allocated to a leaf node as in the case of "kiryu" and "chiryu".

[0062] Next, operation of a speech recognition apparatus in Embodiment 1 of the present invention will be described with reference to a flowchart of Fig. 2.

[0063] In Fig. 2, the speech segment standard pattern 15 is found for each speech segment by learning in advance learned data in which a lot of speakers vocalized. The phoneme merged speech segment standard pattern 16 is found by learning all the learned data of speech segments that are merged. For example, a standard pattern for the speech segment/ {p, t, k, c} i/ can be obtained by learning all the learned data of /pi/, /ti/, /ki/, and /ci/. This is to

be found in advance for each of all the phoneme merged speech segments.

[0064] In the present embodiment, assuming that the appearance probability of a feature parameter vector can be approximated by a sum of a plurality of Gaussian distributions, (which is referred to as a mixed distribution), average value vectors and covariance matrices in the Gaussian distribution are found from the learned data for each frame of the standard pattern.

10 [0065] The phoneme merged speech segment tree 11 is made in advance from the recognition target vocabulary set 10 in development processing of the tree S07.

[0066] First, in acoustic analysis processing S01, an inputted unknown speech signal is converted to D feature parameters for each analysis time (hereinafter referred to as frame). Examples of the feature parameters include LPC cepstrum coefficient and LPC mel-cepstrum coefficient in accordance with linear prediction analysis; mel-LPC cepstrum coefficient in accordance with mel-linear prediction analysis; mel-frequency cepstrum coefficient (MFCC) in accordance with mel-scale filter bank; and the like, and any feature parameter may be used as long as it is suited for speech recognition.

20 [0067] In collation processing S02, in accordance with the phoneme merged speech segment tree 11, a feature

parameter time series for the unknown inputted speech from the acoustic analysis processing S01 is collated with that for the standard pattern while connecting the phoneme merged speech segment standard pattern 16 and the speech segment standard pattern 15. The collation is performed by DP matching using input frame synchronous beam search. Since the method for DP matching and the method for beam search are the same as those of the conventional example, their description is omitted. The collation is referred to as first collation.

[0068] Although the phoneme merged speech segment tree 11 is to be made in advance in the present embodiment, it may also be dynamically developed while performing beam search DP.

[0069] A conceptual diagram as to DP using input frame synchronous beam search is shown in Fig. 9. In Fig. 9, an axis of abscissas represents a frame of inputted speech, while an axis of ordinates represents a frame of a speech segment standard pattern connected in accordance with the phoneme merged speech segment tree. The axis of ordinates serving as a dictionary is in a tree shape. DP matching between the inputted speech and the tree-shaped dictionary is to calculate scores while finding an optimal path of the inputted speech pattern and the standard pattern on the tree-shaped DP screen. In the tree-shaped DP screen,

phonemes from the first to nth steps have been merged, so that the expansion of branches is suppressed.

[0070] In DP matching, pruning DP paths is carried out synchronously with the input frame by beam search. Since
5 the number of candidate lattice points remaining in a beam is far smaller than the number of all the lattice points on the DP screen, the DP screen is not necessary for an actual memory, which is a virtual one.

[0071] After a while from the start of vocalization,
10 cumulative scores of the DP paths in the dictionary, which are not similar to the speech content, become small values enough to be pruned and thus the number of candidate lattice points is drastically reduced. Therefore, the reduction of the number of candidate lattice points so far
15 leads to the reduction of the total calculation quantity. As in the first embodiment, by suppressing the expansion in the vicinity of initial phonemes of words of the tree, the number of candidate lattice points remaining in the beam in the vicinity of the start of speech can greatly be reduced.

[0072] In determination processing S03, a leaf node
20 (maximum likelihood leaf node) is found, and determination of whether or not a word corresponding thereto is uniquely determined is performed. If it is uniquely determined (Y), namely in the case where there is only one word
25 corresponding to the maximum likelihood leaf node, the word

is outputted as a recognition result. If it is not uniquely determined (N), namely in the case where there are a plurality of words corresponding to the maximum likelihood leaf node, the recognition result is determined
5 in the following method.

[0073] In re-collation candidate extraction processing S05, re-collation candidates are selected. In the present embodiment, the re-collation candidates are those vocabularies corresponding to the maximum likelihood leaf
10 node. As another method, there is also a method in which all the vocabularies corresponding not only to the maximum likelihood leaf node but also to the top K leaf nodes of the cumulative scores remaining in the beam.

[0074] Next, in re-collation tree development processing
15 S06, a first to nth steps of a speech segment tree, where phoneme merging is not performed with respect to re-collation candidate vocabularies, are developed. This speech segment tree serves as a first half speech segment re-collation tree 13. In the first half speech segment re-
20 collation tree 13, a recognition vocabulary is uniquely determined in the first to nth steps. Thus, the vocabulary is allocated to an end node of the nth step. An example of the first half speech segment re-collation tree 13, where re-collation candidates are three words of "meguro",
25 "memuro" and "nemuro", i.e., $n=3$, is shown in Fig. 8.

[0075] In the present embodiment, in performing the collation processing S02 by DP matching beforehand, it is required to store an input frame position F_s corresponding to a start node of the first step and an input frame position F_e corresponding to an end node of the n th step.

[0076] In first half re-collation processing S04, an input speech from the frame F_s to the frame F_e is re-collated with a speech segment standard pattern 15 connected in accordance with the first half speech segment re-collation tree 13 by DP matching. In the case of re-collation, since the recognition target vocabulary is small, beam search is not necessarily required to be performed. As a result of the re-collation, a word having the highest cumulative score, which corresponds to the end node of the 10 n th step of the re-collation tree, is outputted as a recognition result.

[0077] In the case of the method in which all the words corresponding to the top K leaf nodes of the cumulative scores remaining in the beam serve as re-collation candidates, a sum S of a score of a first half of 20 vocalization, namely a score S_a from an input frame F_s to a frame F_e , which is found as a result of the re-collation, and a score of a second half of vocalization, namely a score S_b from an input frame F_{e+1} , which is found as a result of the first collation, to an end frame of 25

vocalization is found for each re-collation candidate vocabulary. Then, a word with the largest S is determined as a recognition result.

[0078] In the present embodiment, although phoneme merging is performed collectively from the first to nth steps, it may be performed in all the steps. Further, a portion of the tree, which is densely divided, may be partially modified by further dividing and the like. Instead of collectively performing re-collation to the end node of the nth step, re-collation may be performed to the node in which the word is uniquely determined. In the case where phoneme merging is performed in all the steps, it is not necessary to use the speech segment standard pattern 15 in which phonemes are not merged.

[0079] In the present embodiment, if there is only one words corresponding to the maximum likelihood leaf node, re-collation is not performed. In this case also, not only the maximum likelihood leaf node but also all the words corresponding to the top K leaf nodes of the cumulative scores may serve as re-collation candidates.

[0080] As described above, according to the present embodiment, by using the phoneme merged speech segment tree in which consonants belonging to the same phoneme group are merged as to speech segments from the first to nth steps, the quantity of calculation in the first collation can

greatly be reduced, and, even if re-collation is performed, the total quantity of calculation can greatly be reduced.

[0081] Further, in this method, since similar phonemes are recognized without making any distinction, the possibility that correct candidates are failed to be selected is low. Thus, the quantity of calculation can be reduced without deteriorating recognition performance.

[0082] Furthermore, in the present embodiment, storing the input frame position F_s corresponding to the start node of the first step and the input frame position F_e corresponding to the end node of the n th step, re-collation is performed only between the F_s and the F_e , so that the quantity of calculation required for re-collation is so small.

[0083] (Embodiment 2)

Next, operation of a speech recognition apparatus according to Embodiment 2 of the present invention will be described with reference to a flowchart of Fig. 10.

[0084] What differs from Embodiment 1 is that the first half speech segment re-collation tree 13 and the first half re-collation processing S04 are changed to a speech segment re-collation tree 14 and re-collation processing S21, respectively. Different from Embodiment 1, the re-collation speech segment tree 14 is a tree representing not only from a first to n th steps but also to an end step of a

word.

[0085] Since operation of Embodiment 2 is almost the same as that of Embodiment 1, only different parts will be described.

5 [0086] In Embodiment 1, re-collation is performed only on the input section corresponding to the first to nth steps, where phoneme merging of the phoneme merged speech segment tree was performed in the first collation. On the other hand, in Embodiment 2, re-collation is performed on
10 all the vocalization section.

[0087] In re-collation tree development processing S09, a speech segment tree, where phoneme merging is not performed on re-collation candidate words, is developed. This speech segment tree is referred to as the re-collation
15 speech segment tree 14. The re-collation speech segment tree 14 is a speech segment tree representing not from the first to nth steps but to the end step of the word.

[0088] An example of the re-collation speech segment tree, where re-collation candidates are three words, i.e.,
20 "meguro", "memuro" and "nemuro", will be described in Fig. 11.

[0089] In the present embodiment, in collation processing S02, it is not required to store an input frame position corresponding to a start node of the first step
25 and an input frame position corresponding to an end node of

the nth step.

[0090] In re-collation processing S21, an input speech from the start of vocalization to the end of vocalization is re-collated with a speech segment standard pattern 15 connected in accordance with the re-collation speech segment tree 14 by DP matching. In the case of re-collation, since the recognition target vocabulary is small as in Embodiment 1, it is not necessarily required to perform beam search.

[0091] As a result of the re-collation processing S21, a vocabulary corresponding to a leaf node of the re-collation tree, which had the highest cumulative score, is outputted as a recognition result.

[0092] As described above, according to Embodiment 2, in the case where the input frame position corresponding to the end node of the nth step was not an optimal position in the first collation, by performing re-collation from the start to the end of the vocalization section, more accurate collation can be performed. Thus, recognition performance is further improved compared with Embodiment 1.

[0093] In Embodiment 2, since it is not required to store the input frame position F_s corresponding to the start node of the first step and the input frame position F_e corresponding to the end node of the nth step, the first recognition processing and memory capacity are less than

those of Embodiment 1.

[0094] As in Embodiment 2, in the case where the vocalization section is re-collated from its start to end, a distance scale completely different from the one used for the first collation may be used. Therefore, in performing re-collation, higher recognition performance can also be achieved using a method by which higher recognition can be achieved with only several words.

[0095] (Embodiment 3)

10 Next, operation of a speech recognition apparatus according to Embodiment 3 of the present invention will be described with reference to a flowchart of Fig. 12

[0096] What differs from Embodiment 1 is that the phoneme merged speech segment tree 11, the phoneme merged speech segment standard pattern 16, the speech segment standard pattern 15 are changed to a rough speech segment tree 12, a rough speech segment standard pattern 17, an accurate speech segment standard pattern 18, respectively, and that the determination processing S03 is not required.

20 [0097] The accurate speech segment standard pattern 18 is the same as the speech segment standard pattern 15 of Embodiment 1. In Embodiment 3, a normal speech segment is referred to as an accurate speech segment in order to contrast it with a rough speech segment.

25 [0098] The rough speech segment tree 12 and the rough

speech segment standard pattern 17 will hereinafter be described. The rough speech segment is defined as the one that is obtained by reducing the accuracy of the speech segment standard pattern. The following two methods are conceived as a method therefor.

[0099] The first is a method in which the quantity of distance calculation required per one rough speech segment is reduced, compared with that required per one accurate speech segment. Specifically, a method for reducing the number of frames of the rough speech segment pattern, a method for reducing the mixture number of the Gaussian distribution, a method for making the covariance matrices of the Gaussian distribution common to reduce the number of kinds of covariance matrices, and the like are conceived.

In this method, the shape of the speech segment tree is not changed.

[0100] The second is a method for merging speech segments of different phonological environment in a range in which the recognition result is uniquely determined. By this method also, since arcs and nodes of the tree are reduced, the quantity of calculation can be reduced. For example, the case where VCs, whose vowel portions are the same, are merged with one speech segment even if their subsequent consonants are different and the like are conceived. In this method, the shape of the speech segment

tree may be changed. In the case where phonemes are used as units of the speech segment, speech segments are determined to be different in many cases depending on phonemic environment around them, but in the case where they have the same center phoneme, they are merged with one speech segment, whereby the expansion of the tree can greatly be suppressed. As a matter of course, if they have the same center phoneme, the recognition result is naturally and uniquely determined even if merging is performed.

[0101] In Embodiment 3, both of the method for reducing the number of frames of the standard pattern and the method in which VCs having the same vowel portion and different subsequent consonants are merged are used. The former is represented by affixing a bar on a speech segment symbol, while the latter is represented by replacing a consonant portion with an asterisk.

[0102] Fig. 13 is a rough speech segment tree obtained by converting speech segments from the first to third steps ($n=3$) of the base speech segment tree of Fig. 3 to rough speech segments. Speech segments after the fourth step are the same as those of the base speech segment tree. The shape of the rough speech segment tree is slightly different from Fig. 2. VC-merging may also be limited to the only case where the subsequent consonant belongs to the

same phoneme group.

[0103] The rough speech segment standard pattern 17 is provided in advance by learning. Based on the number of frames of the standard pattern, learning is performed by
5 reducing the number of frames of the standard pattern to a half of the number of original frames. Moreover, as to VCs, learning is performed from all the learned data of the speech segments having the same vowel portion. For example, the standard pattern for the speech segment /e*/ can be
10 obtained by learning all the learned data of the speech segments whose vowel portions are /e/ and the subsequent consonants are different, i.e., /em/, /en/, /eg/, /eb/,

[0104] Operation of Embodiment 3 is almost the same as that of Embodiment 1 and thus only different portions will
15 be described.

[0105] In collating processing S02, in the same manner as in Embodiment 1, a feature parameter time series of an unknown inputted speech is collated with that of the standard pattern while connecting the rough speech segment
20 standard pattern 17 and the accurate speech segment standard pattern 18 in accordance with the rough speech segment tree 12.

[0106] After performing the collation, re-collation candidates are extracted in re-collation candidate
25 extraction processing S05. In the present embodiment, K

vocabularies corresponding to the top K leaf nodes of the cumulative scores serve as re-collation candidates. In the same manner as in Embodiment 1, the first half speech segment re-collation tree 13 is developed for each re-collation candidate, and a first half of vocalization is collated with the accurate speech segment standard pattern in the first half re-collation processing S04.

[0107] A sum S of a score Sa of a first half of vocalization, which is found as a result of the re-collation, and a score Sb of a second half of vocalization, which is found as a result of the first re-collation, is found for all the candidate vocabularies. Then, a vocabulary with the largest S is determined as a recognition result.

[0108] In the present embodiment, although phoneme merging is performed collectively from the first to nth steps, a portion of a tree, which is densely divided, may be partially modified by further dividing and the like. Instead of collectively performing re-collation to the end node of the nth step, re-collation may be performed until the node in which the word is uniquely determined.

[0109] As described above, according to Embodiment 3, by using the rough speech segment tree whose accuracy of the speech segment standard pattern is reduced, the calculation quantity required for the re-collation of the rough speech

segments is small and thus the calculation quantity in the first collation can greatly be reduced. Even if recollation is performed, the total calculation quantity can be reduced.

- 5 [0110] Since rough collation is performed in a part where the calculation quantity is large, while accurate collation is performed in a part where the calculation quantity is small a little while after the start of vocalization, the efficiency is high.